Supplementary text and tables for "Integrative modelling of gene and genome evolution roots the archaeal tree of life" Williams et al.

Table S1: 45 archaeal phylogenomic markers used in this study, indicating the overlap with the dataset of Petitjean et al. (1).

OMA ID (this study)	ArCOG	COG	COG category	Used in (1)?
OG410	ArCOG00042	COG0037	D	
OG415	ArCOG01563	COG5257	J	
OG280	ArCOG04169	COG0201	U	
OG1439	ArCOG04087	COG0098	J	Y
OG3774	ArCOG04099	COG0185	J	Y
OG2564	ArCOG04239	COG0522	J	Y
OG2721	ArCOG04092	COG0094	J	Y
OG2160	ArCOG04176	COG1976	J	
OG1219	ArCOG04245	COG0052	J	
OG426	ArCOG01742	COG1503	J	
OG1457	ArCOG04071	COG0088	J	Y
OG4255	ArCOG01758	COG0051	J	Y
OG1606	ArCOG00675	COG1095	K	
OG3506	ArCOG04091	COG0096	J	Y
OG558	ArCOG01001	COG0024	J	
OG3341	ArCOG04229	COG1781	F	
OG1896	ArCOG04289	COG0081	J	Y
OG3779	ArCOG04089	COG2147	J	
OG73	ArCOG00543	COG1782	R	
OG4714	ArCOG04096	COG0186	J	Y
OG354	ArCOG00982	COG0371	C	
OG673	ArCOG04288	COG0244	J	Y
OG3259	ArCOG04185	COG0184	J	Y
OG494	ArCOG00357	COG0012	J	
OG792	ArCOG01358	COG0621	J	Y
OG46	ArCOG00187	COG1245	R	

OG329	ArCOG04050	COG0258	L	
OG42	ArCOG01559	COG0480	J	
OG1514	ArCOG04067	COG0090	J	Y
OG769	ArCOG04174	COG0731	J	
OG3189	ArCOG04095	COG0093	J	Y
OG1194	ArCOG04107	COG1093	J	
OG2277	ArCOG04098	COG0091	J	Y
OG1107	ArCOG04187	COG1500	J	
OG1076	ArCOG04070	COG0087	J	Y
OG2892	ArCOG01344	COG2238	J	
OG201	ArCOG00412	COG0072	J	
OG3090	ArCOG04240	COG0100	J	Y
OG450	ArCOG01228	COG0541	U	Y
OG2240	ArCOG01722	COG0099	J	Y
OG5673	ArCOG04287	COG2058	J	
OG2679	ArCOG04255	COG0048	J	Y
OG4370	ArCOG04228	COG1990	J	
OG260	ArCOG01561	COG5256	J	
OG2258	ArCOG04254	COG0049	J	Y

Table S2: 29 universally-conserved genes used for rooting with a bacterial outgroup. See also (2, 3).

Gene ID (*Saccharomyces cerevisiae)*
Rps14bp
Rps23bp
Fun12p
Rpl11ap
Rps3p
Rps16ap
Rpl1ap
Rpl2bp
Rpl23bp
Rpl12ap
Eft1p
Kae1p
Rps0bp

Rps5p
Rps2p
Srp54p
Tef1p
Rli1p
Dps1p
Rpa190p
Sec61p
Cct5p
Rfc2p
Vma2p
Map2p
Rpl16ap
Gln4p
Rpa135p
Srp101p

Table S3: Approximately-unbiased test for the archaeal root, with DPANN Archaea included in the analysis. "Th" refers to the position of the *Thermococcales*. "Basal" implies that the root lies between this group and all other Archaea. Bold, underlined roots are those which could not be rejected by the analysis at P > 0.05 (i.e., AU > 0.05).

Root	**ΔlnL**	**AU**
DPANN basal, Th with Eury	**-263.7**	**1**
DPANN basal, TACKL+Th	263.7	2×10^{-74}
Lokiarchaeum basal	504.7	5×10^{-6}
TACKL basal	538.5	4×10^{-44}
TACKL+Th basal	692.6	2×10^{-52}
Euryarchaeota basal	786.3	2×10^{-54}
Raymann et al.	1025.6	7×10^{-80}

Table S4: Approximately-unbiased test for the archaeal root, without the DPANN Archaea. Bold, underlined roots are those which could not be rejected by the analysis at P > 0.05 (i.e., AU > 0.05).

Root	**ΔlnL**	**AU**
***Lokiarchaeum* basal**	-9.9	**0.631**
TACKL basal	9.9	**0.369**
TACKL+Th basal	374.1	2×10^{-5}

Raymann et al.	448.2	2×10^{-5}

Table S5: Approximately-unbiased test for the archaeal root, considering only gene families containing at least one sequence from the DPANN Archaea. Bold, underlined roots are those which could not be rejected by the analysis at P > 0.05 (i.e., AU > 0.05).

Root	**ΔlnL**	**AU**
DPANN basal, (TACKL+Th)	**-26.9**	**0.906**
DPANN basal, Th with Eury	**26.9**	**0.157**
Lokiarchaeum basal	103.2	0.001
TACKL basal	170.0	7×10^{-53}
TACKL+Th basal	273.4	2×10^{-88}
Euryarchaeota basal	358.0	1×10^{-51}
Raymann et al.	405.7	2×10^{-41}

Table S6: Annotations for the archaeal gene families analysed in our study, and mapping to key nodes on the tree. The annotation was based upon ArCOG (4); families were mapped to a node when the probability of at least one gene copy was >= 0.5. This table has been deposited at https://doi.org/10.6084/m9.figshare.4657396.v2

		Minimum estimate	5% Minimum estimate	Median estimate	95% Maximum estimate	Maximum estimate
Full alignment	Ancestor of DPANN	63.9	64.9	70.6	76.8	78.7
	Ancestor of Core Euryarchaeota	63.4	64.5	70.4	76.1	76.5
	Ancestor of TACK+*Lokiarchaeum*	70.0	71.1	77.4	84.2	85.3
	Ancestor of TACK+*Lokiarchaeum*/Euryarchaeota	65.4	66.8	73.1	79.0	79.9
	Ancestor of all Archaea	66.0	66.7	73.1	78.9	80.5

Fewer gaps (Gap 15)	Ancestor of DPANN	59.8	61.9	71.1	80.5	82.3
	Ancestor of Core Euryarchaeota	57.4	57.9	67.1	76.0	77.6
	Ancestor of TACK/*Lokiarchaeum*	63.9	67.1	76.9	87.2	90.7
	Ancestor of TACK+*Lokiarchaeum*/Euryarchaeota	58.5	61.8	71.3	79.6	83.4
	Ancestor of all Archaea	58.1	61.9	71.2	80.5	82.3
No DPANN	Ancestor of Core Euryarchaeota	67.7	69.1	74.5	80.3	82.0
	Ancestor of TACK+*Lokiarchaeum*	73.2	74.3	80.6	86.9	87.8
	Ancestor of TACK+*Lokiarchaeum*/Euryarchaeota	69.4	70.2	75.7	81.8	82.9

Table S7: Estimates of optimal growth temperatures (OGT) for select ancestors in the Archaeal tree, based on three different data sets. For each data set a linear regression was computed between the second axis of a correspondence analysis on amino acid content and optimal growth temperature. Estimates are based on 100 sequences sampled according to site-wise probabilities computed under the LG+4G+CoaLa model (5) with 3 axes. In addition to the median OGT predicted for each sample of 100 sequences, we provide minimum, maximum, and 5% and 95% quantile estimates to show the spread of the estimated OGTs. These values take into account the variance across the 100 sampled sequences as well as the uncertainty in the parameters of the linear regression used to predict OGT.

Table S8: Inferred branch-wise numbers of gene originations, duplications, transfers and losses mapped onto the rooted archaeal species tree. These values were obtained from the

maximum-likelihood rooted tree inferred from the entire dataset. Extant taxa are denoted by an abbreviated species name, while interior branches are indicated by a number. The mapping of numbers to internal branches is given by Figure S21. Gene acquisitions are the sum of originations and transfers-in; transfers-out gives the number of genes inferred to have been donated from a given branch to other branches on the tree. This table has been deposited at https://doi.org/10.6084/m9.figshare.4657396.v2

Tables S9-S10: Annotations for the gene families inferred to be acquired or expanded on the haloarchaeal (S9) and thaumarchaeotal (S10) stems. These tables have been deposited at https://doi.org/10.6084/m9.figshare.4657396.v2

Supplementary Figures

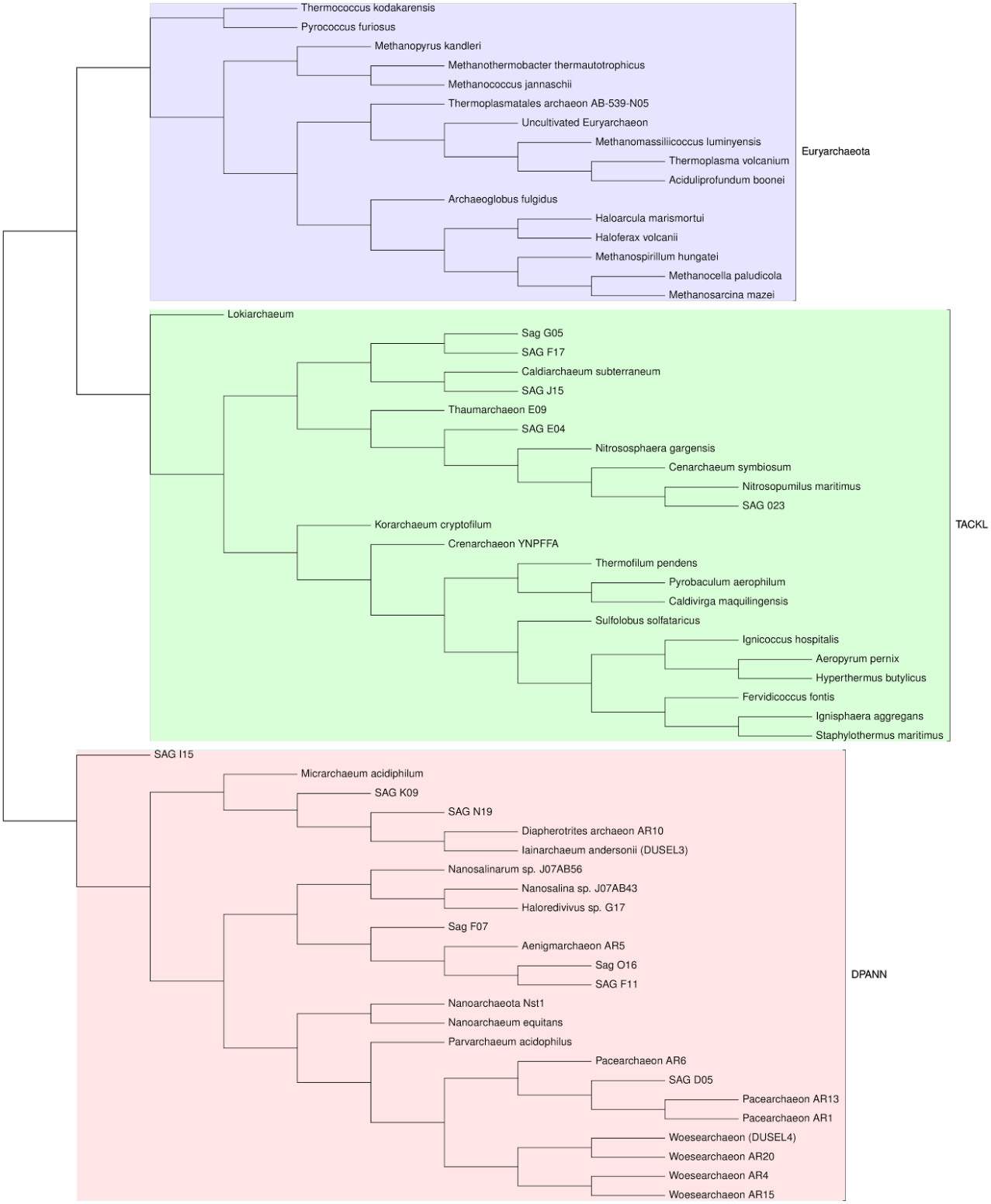


Figure S1: Unrooted matrix representation parsimony (6, 7) supertree of the Archaea based on 3,242 single gene trees. Input majority-rule posterior consensus single gene trees were inferred under the C60+LG model in PhyloBayes (8); branches with less than 0.5 posterior probability were

collapsed. The unrooted topology is closely similar to that obtained by an analysis of 45 concatenated protein markers (Figure 1) under the CAT+GTR model. The difference relates to the placement of the *Thermococcales,* which group at the base of the Euryarchaeota (this tree) or the base of the TACKL lineage (Figure 1, concatenated protein tree).

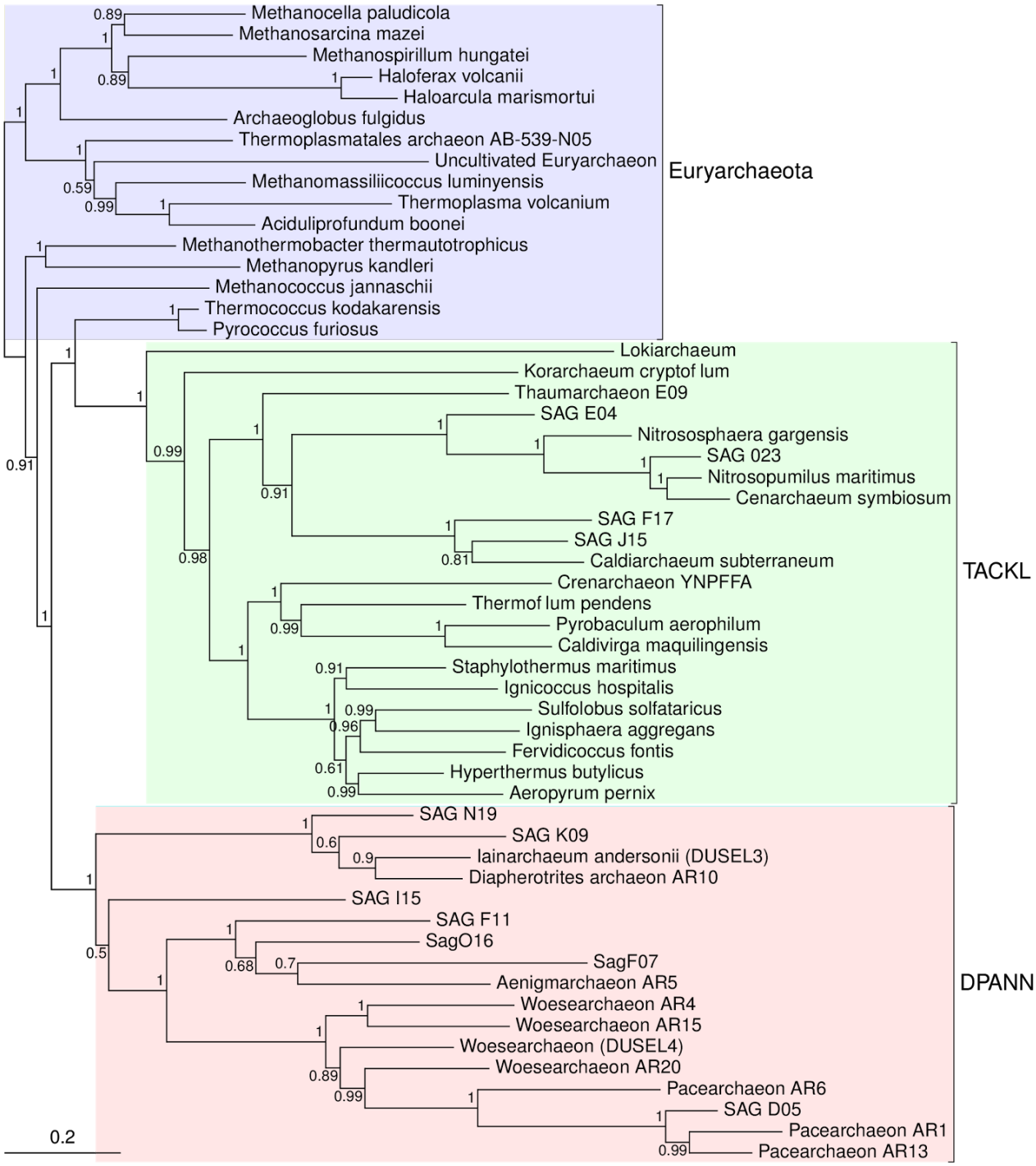


Figure S2: Re-analysis of the 45 gene concatenated protein alignment in which the longest-branching DPANN lineages have been removed. DPANN clanhood is obtained with

maximal posterior support (PP = 1). Support values are Bayesian posterior probabilities, and the tree is rooted according to the maximum likelihood root position obtained in the DTL analysis.

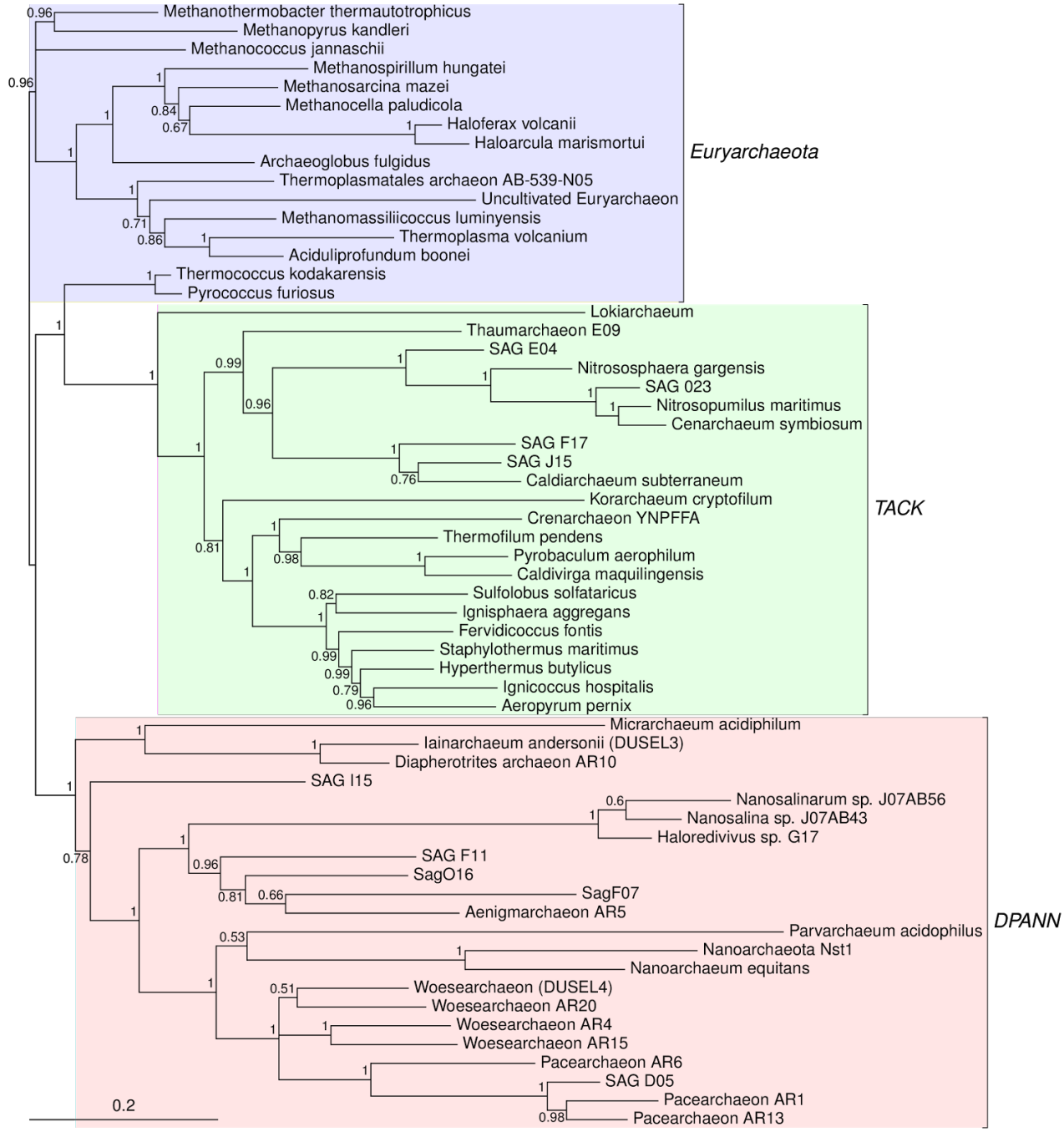


Figure S3: Re-analysis of the 45 gene concatenated protein alignment in which the alignment has been masked using the BLOSUM62 setting in BMGE (9). This matrix had 5,920 amino acid positions, in contrast to the 10,738 positions used in the main analysis. DPANN clanhood is obtained with maximal posterior support (PP = 1). Support values are Bayesian posterior probabilities, and the tree is rooted according to the maximum likelihood root position obtained in

the DTL analysis.

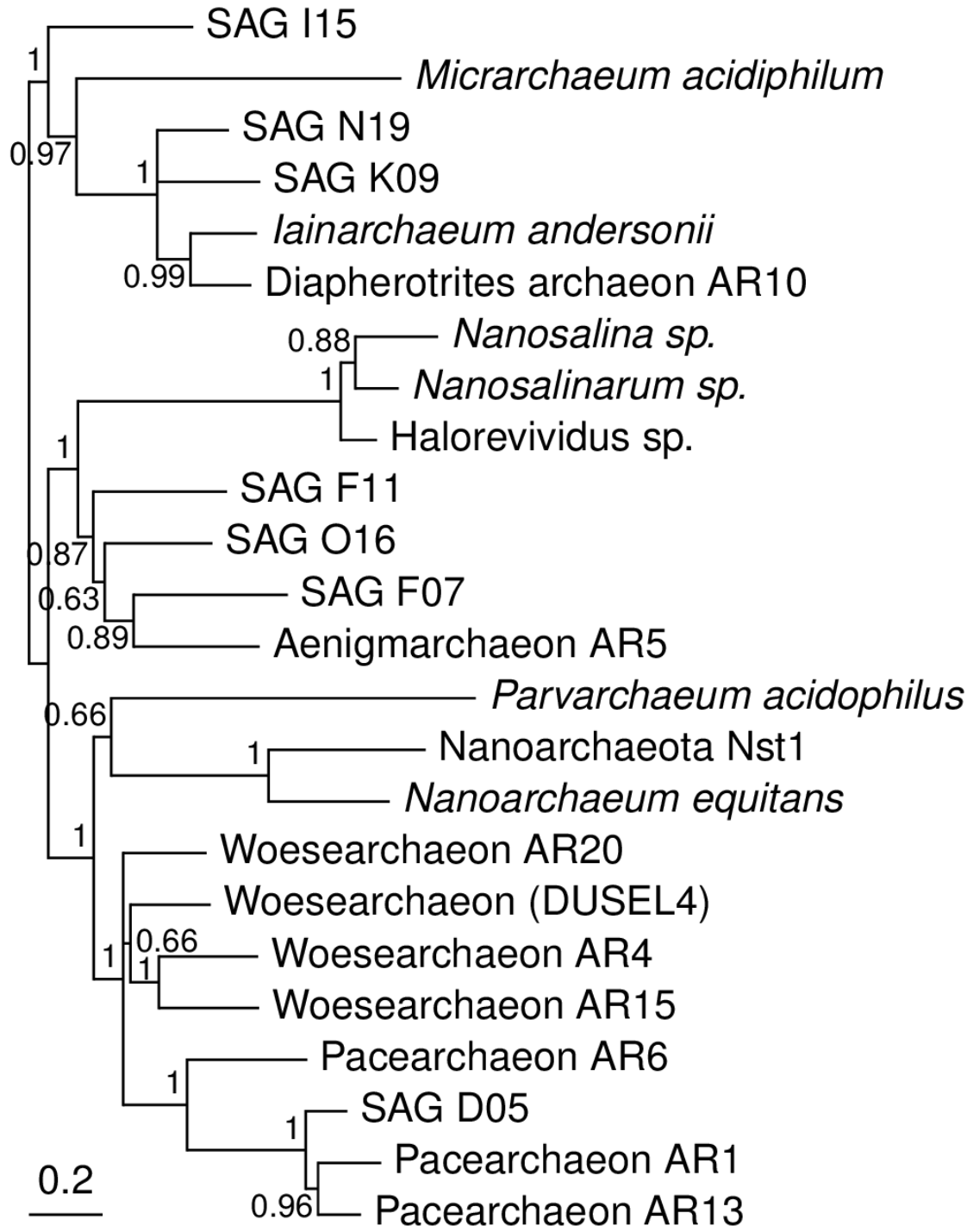


Figure S4: Re-analysis of the 45 gene concatenated protein alignment including only the DPANN lineages. The relationships among DPANN lineages are entirely consistent with those inferred from the entire dataset, providing no evidence that DPANN clanhood was an artifact of LBA in that analysis. Support values are Bayesian posterior probabilities, and the tree is rooted according to Figure 1.

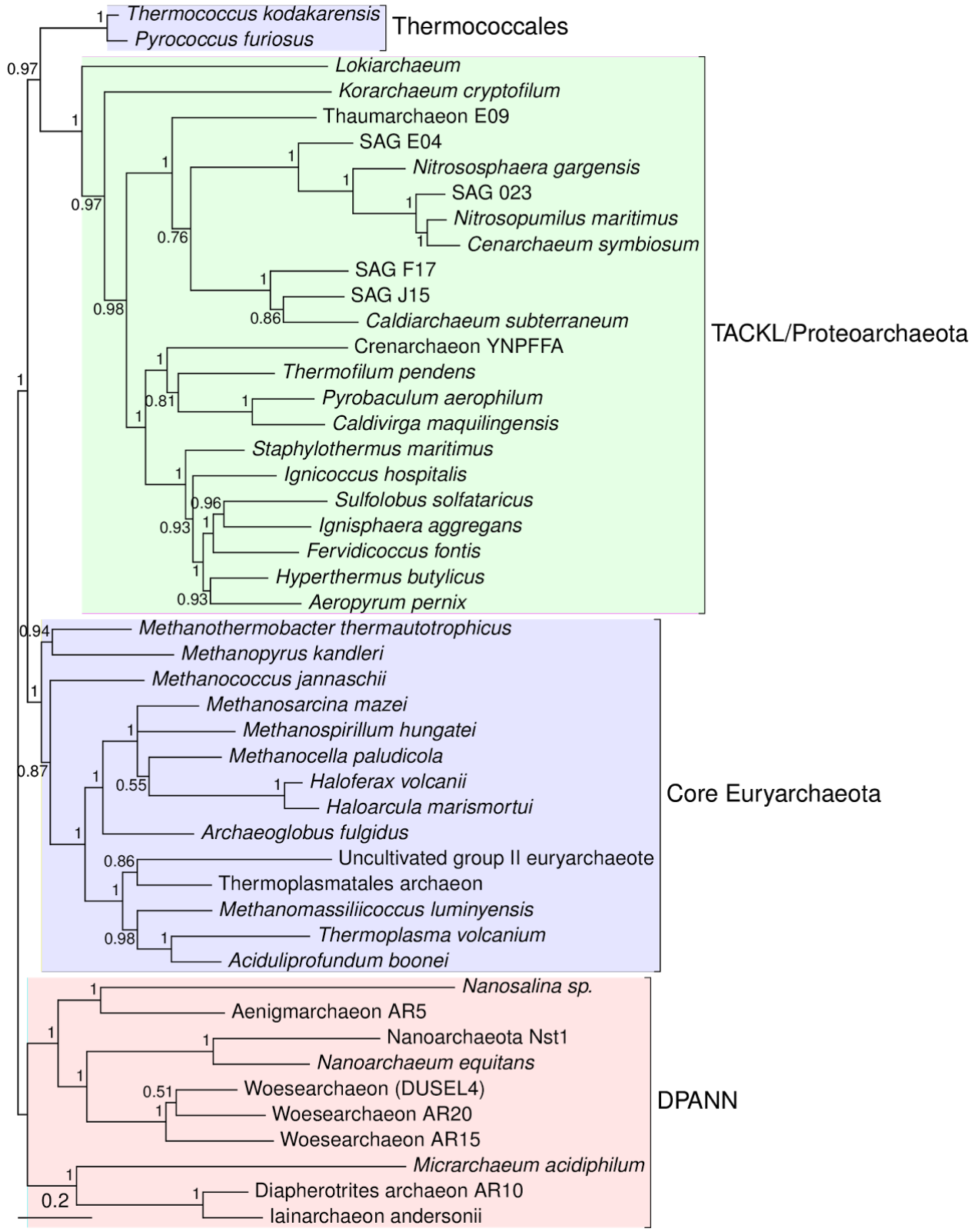


Figure S5: Analysis of a subsampled protein concatenation in which DPANN gene representation is equal to that of the other archaeal clans. The matrix was obtained by including the ten most complete DPANN genomes (including at least one from each DPANN sub-lineage)

and the 25 most well-represented genes. The topology inferred under CAT+GTR from a four-state Dayhoff-recoded alignment is very similar to that obtained in Figure 1, and is identical within the DPANN. Support values are Bayesian posterior probabilities, and the tree is rooted according to the maximum likelihood root position obtained in the DTL analysis.

Figures S6-S12: Analyses of subsampled protein concatenations in which only one DPANN lineage was included. S6: Diapherotrites; S7: Parvarchaeum; S8: Aenigmaarchaeota; S9: Nanoarchaeota; S10: Nanohaloarchaeota; S11: Pacearchaeota; S12: Woesearchaeota. The topologies were inferred under CAT+GTR from a four-state Dayhoff-recoded alignment. Support values are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.

Figure S6:

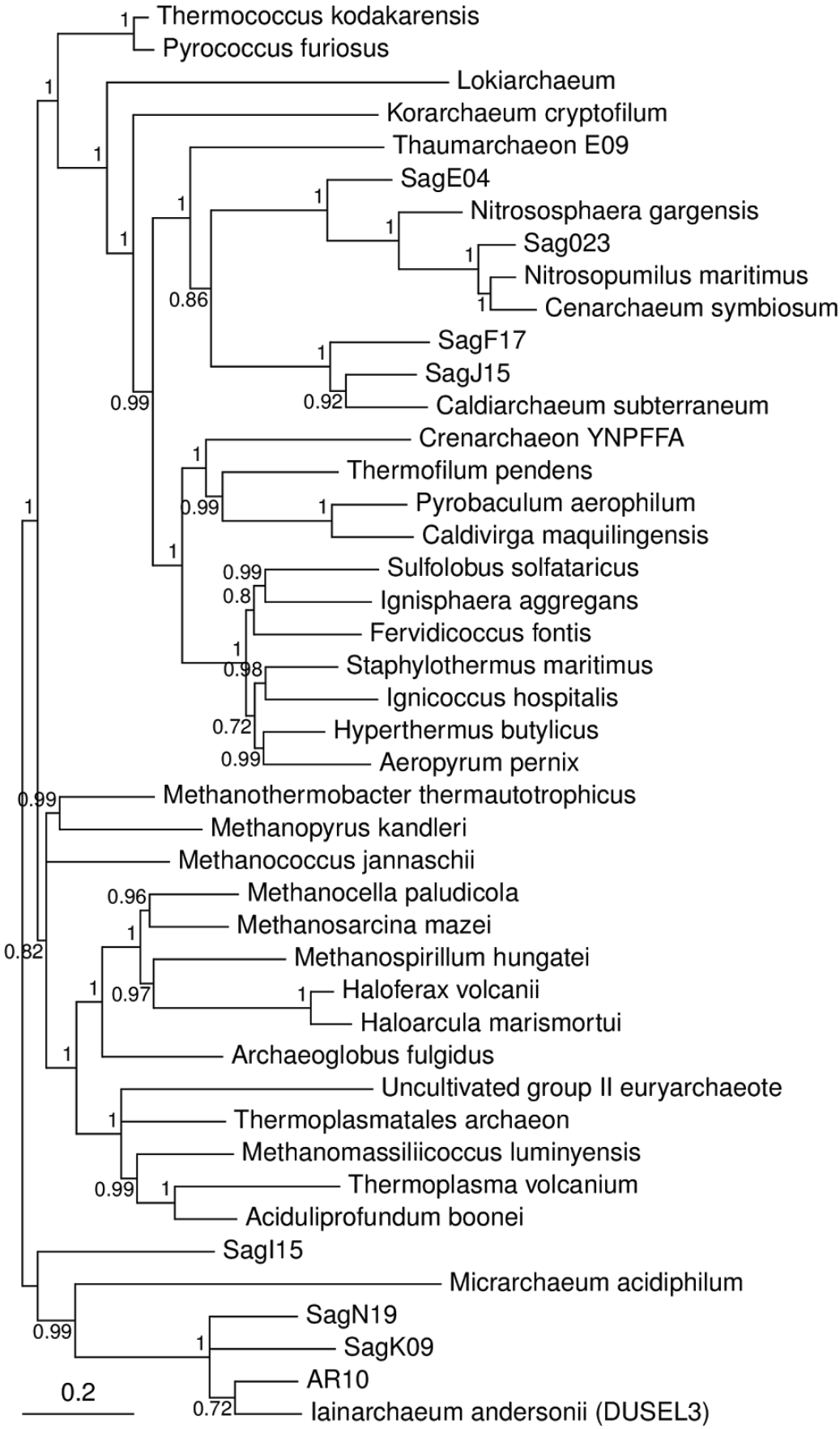

Thermococcus kodakarensis
Pyrococcus furiosus
Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Staphylothermus maritimus
Ignicoccus hospitalis
Hyperthermus butylicus
Aeropyrum pernix
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Methanococcus jannaschii
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Uncultivated group II euryarchaeote
Thermoplasmatales archaeon
Methanomassiliicoccus luminyensis
Thermoplasma volcanium
Aciduliprofundum boonei
SagI15
Micrarchaeum acidiphilum
SagN19
SagK09
AR10
Iainarchaeum andersonii (DUSEL3)
0.2

Figure S7:

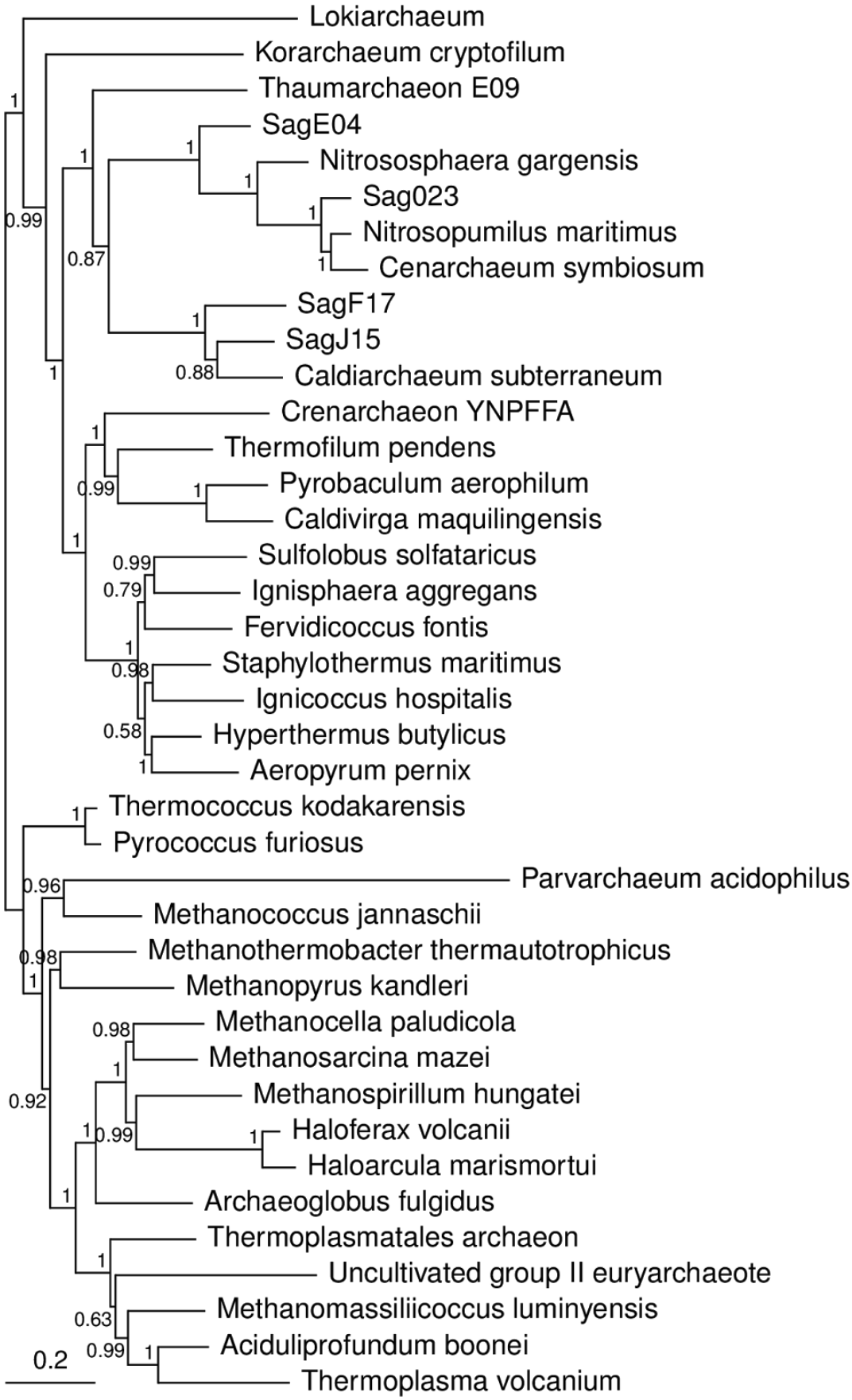

Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Staphylothermus maritimus
Ignicoccus hospitalis
Hyperthermus butylicus
Aeropyrum pernix
Thermococcus kodakarensis
Pyrococcus furiosus
Parvarchaeum acidophilus
Methanococcus jannaschii
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Thermoplasmatales archaeon
Uncultivated group II euryarchaeote
Methanomassiliicoccus luminyensis
Aciduliprofundum boonei
Thermoplasma volcanium
0.2

Figure S8:

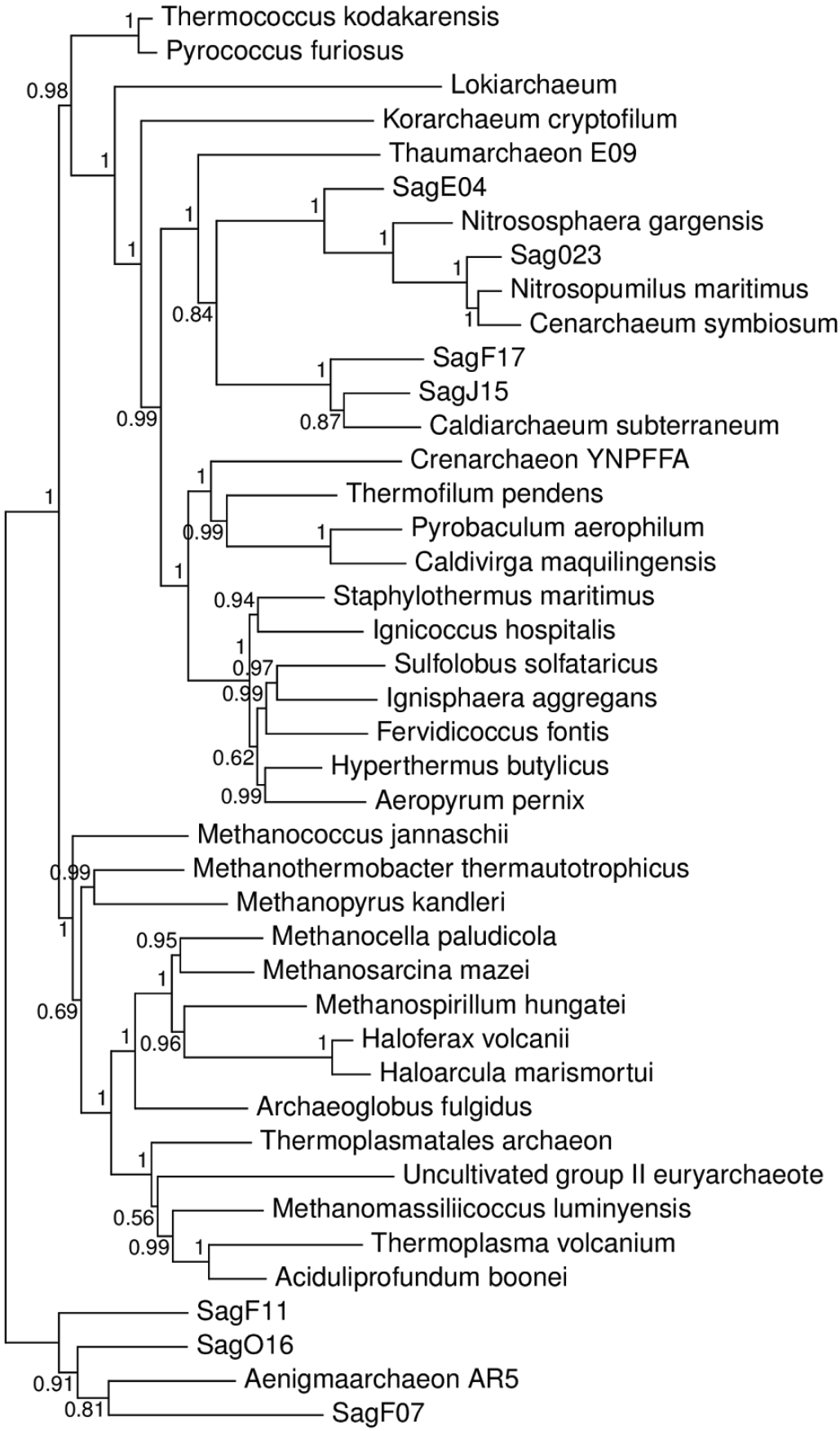

Thermococcus kodakarensis
Pyrococcus furiosus
Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Staphylothermus maritimus
Ignicoccus hospitalis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Hyperthermus butylicus
Aeropyrum pernix
Methanococcus jannaschii
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Thermoplasmatales archaeon
Uncultivated group II euryarchaeote
Methanomassiliicoccus luminyensis
Thermoplasma volcanium
Aciduliprofundum boonei
SagF11
SagO16
Aenigmaarchaeon AR5
SagF07

Figure S9:

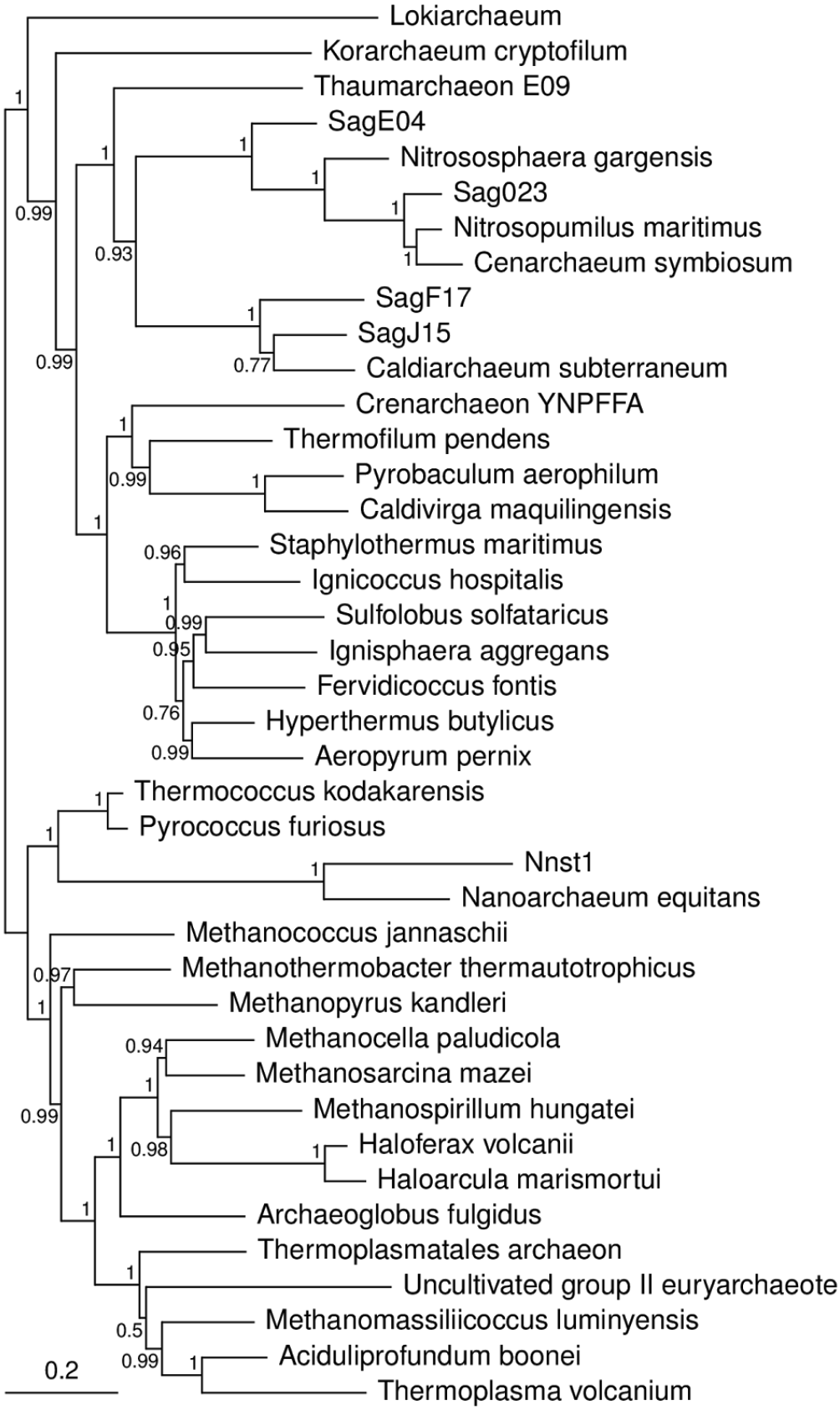

Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Staphylothermus maritimus
Ignicoccus hospitalis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Hyperthermus butylicus
Aeropyrum pernix
Thermococcus kodakarensis
Pyrococcus furiosus
Nnst1
Nanoarchaeum equitans
Methanococcus jannaschii
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Thermoplasmatales archaeon
Uncultivated group II euryarchaeote
Methanomassiliicoccus luminyensis
Aciduliprofundum boonei
Thermoplasma volcanium
0.2

Figure S10:

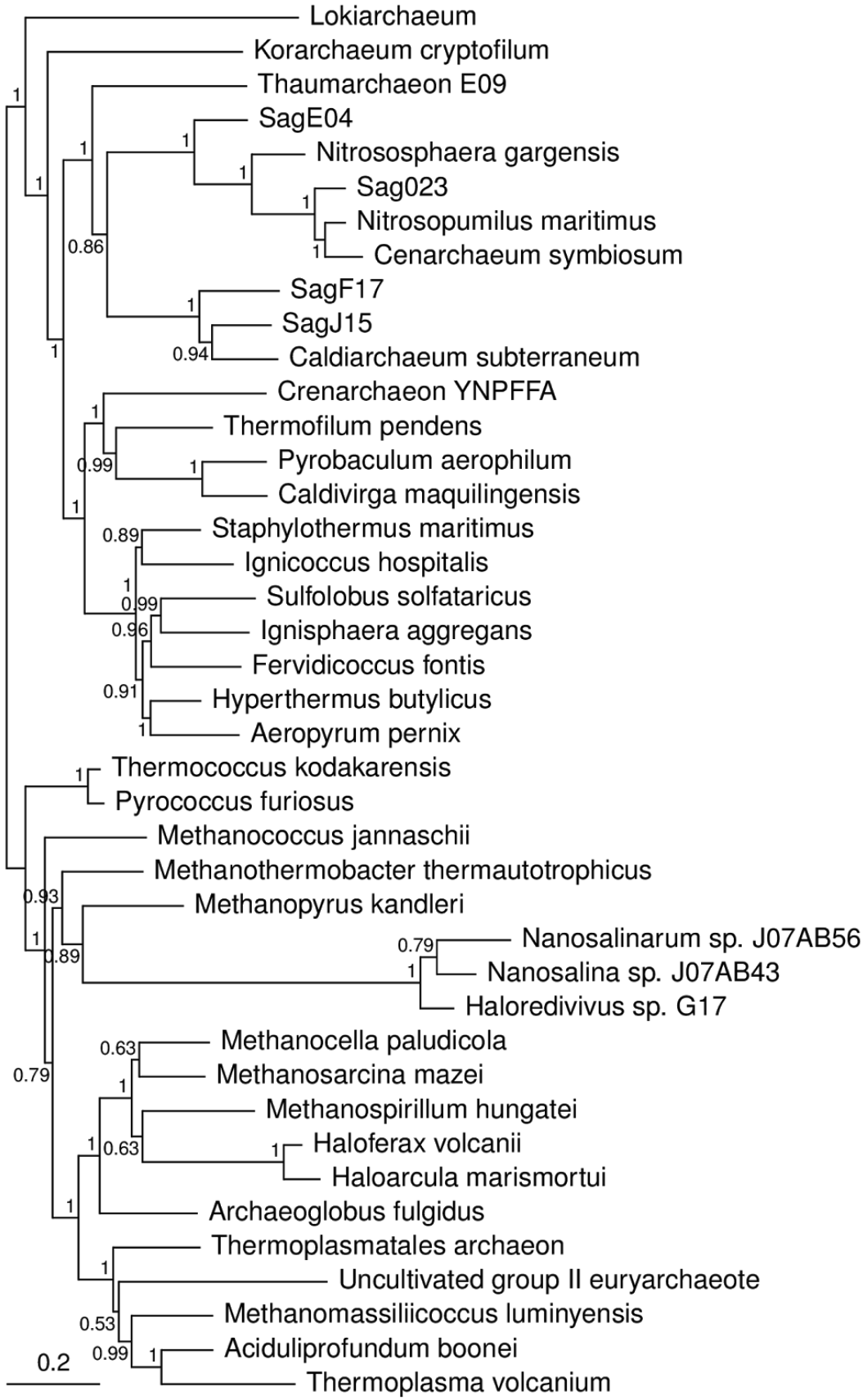

Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Staphylothermus maritimus
Ignicoccus hospitalis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Hyperthermus butylicus
Aeropyrum pernix
Thermococcus kodakarensis
Pyrococcus furiosus
Methanococcus jannaschii
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Nanosalinarum sp. J07AB56
Nanosalina sp. J07AB43
Haloredivivus sp. G17
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Thermoplasmatales archaeon
Uncultivated group II euryarchaeote
Methanomassiliicoccus luminyensis
Aciduliprofundum boonei
Thermoplasma volcanium
0.2

Figure S11:

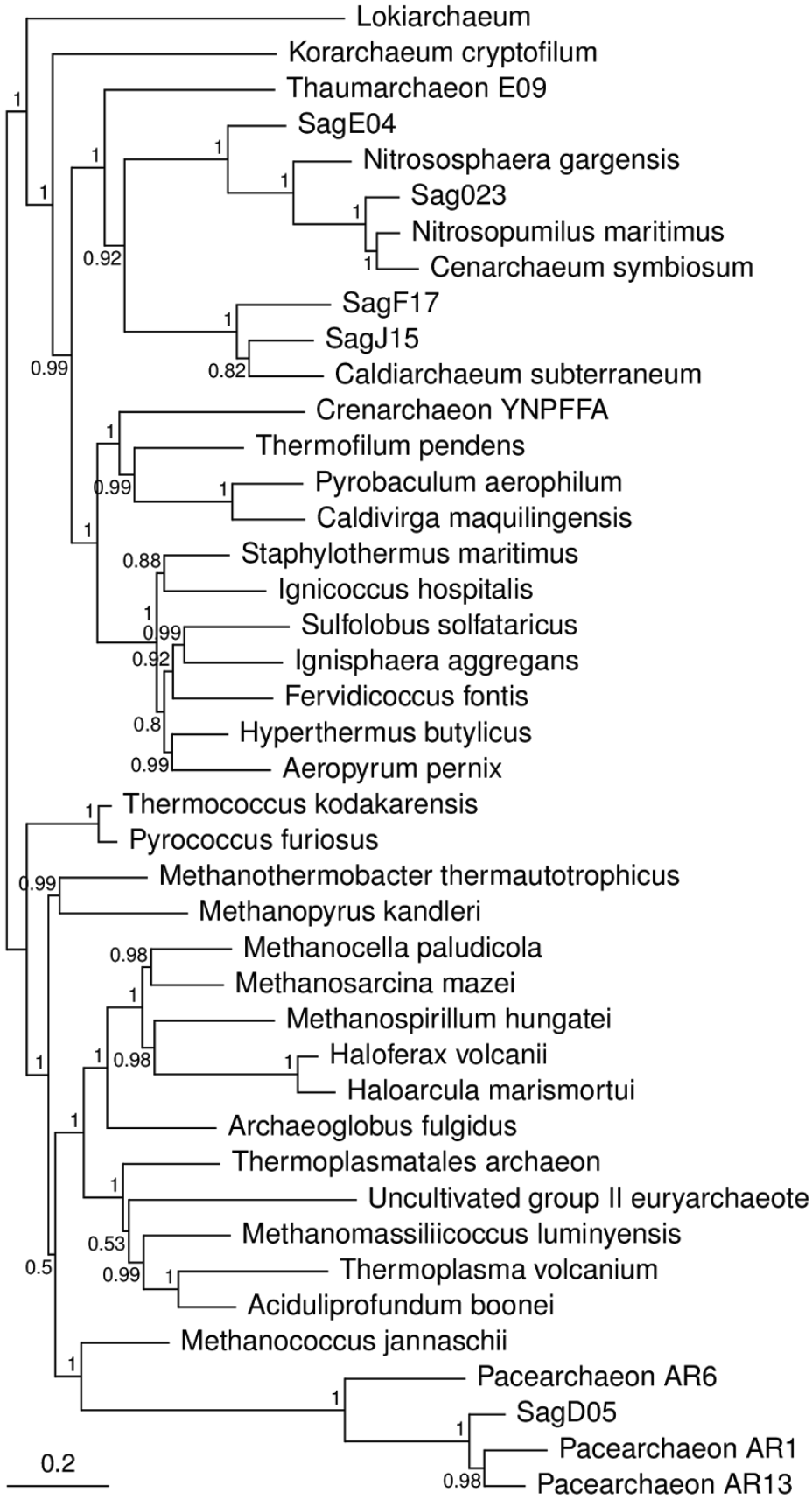
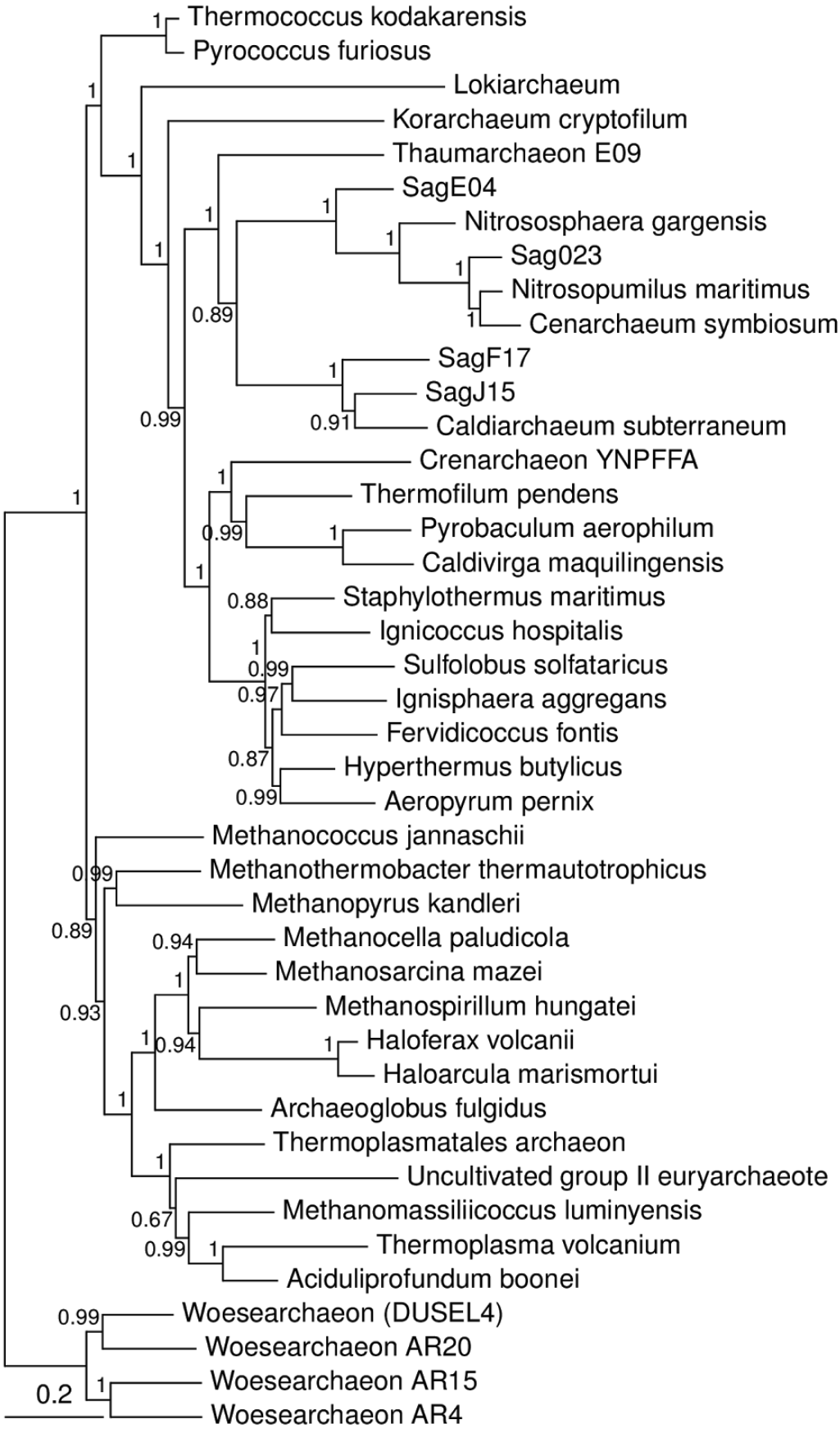

Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Staphylothermus maritimus
Ignicoccus hospitalis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Hyperthermus butylicus
Aeropyrum pernix
Thermococcus kodakarensis
Pyrococcus furiosus
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Thermoplasmatales archaeon
Uncultivated group II euryarchaeote
Methanomassiliicoccus luminyensis
Thermoplasma volcanium
Aciduliprofundum boonei
Methanococcus jannaschii
Pacearchaeon AR6
SagD05
Pacearchaeon AR1
Pacearchaeon AR13
0.2

Figure S12:


Thermococcus kodakarensis
Pyrococcus furiosus
Lokiarchaeum
Korarchaeum cryptofilum
Thaumarchaeon E09
SagE04
Nitrososphaera gargensis
Sag023
Nitrosopumilus maritimus
Cenarchaeum symbiosum
SagF17
SagJ15
Caldiarchaeum subterraneum
Crenarchaeon YNPFFA
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Staphylothermus maritimus
Ignicoccus hospitalis
Sulfolobus solfataricus
Ignisphaera aggregans
Fervidicoccus fontis
Hyperthermus butylicus
Aeropyrum pernix
Methanococcus jannaschii
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Haloferax volcanii
Haloarcula marismortui
Archaeoglobus fulgidus
Thermoplasmatales archaeon
Uncultivated group II euryarchaeote
Methanomassiliicoccus luminyensis
Thermoplasma volcanium
Aciduliprofundum boonei
Woesearchaeon (DUSEL4)
Woesearchaeon AR20
Woesearchaeon AR15
Woesearchaeon AR4
0.2

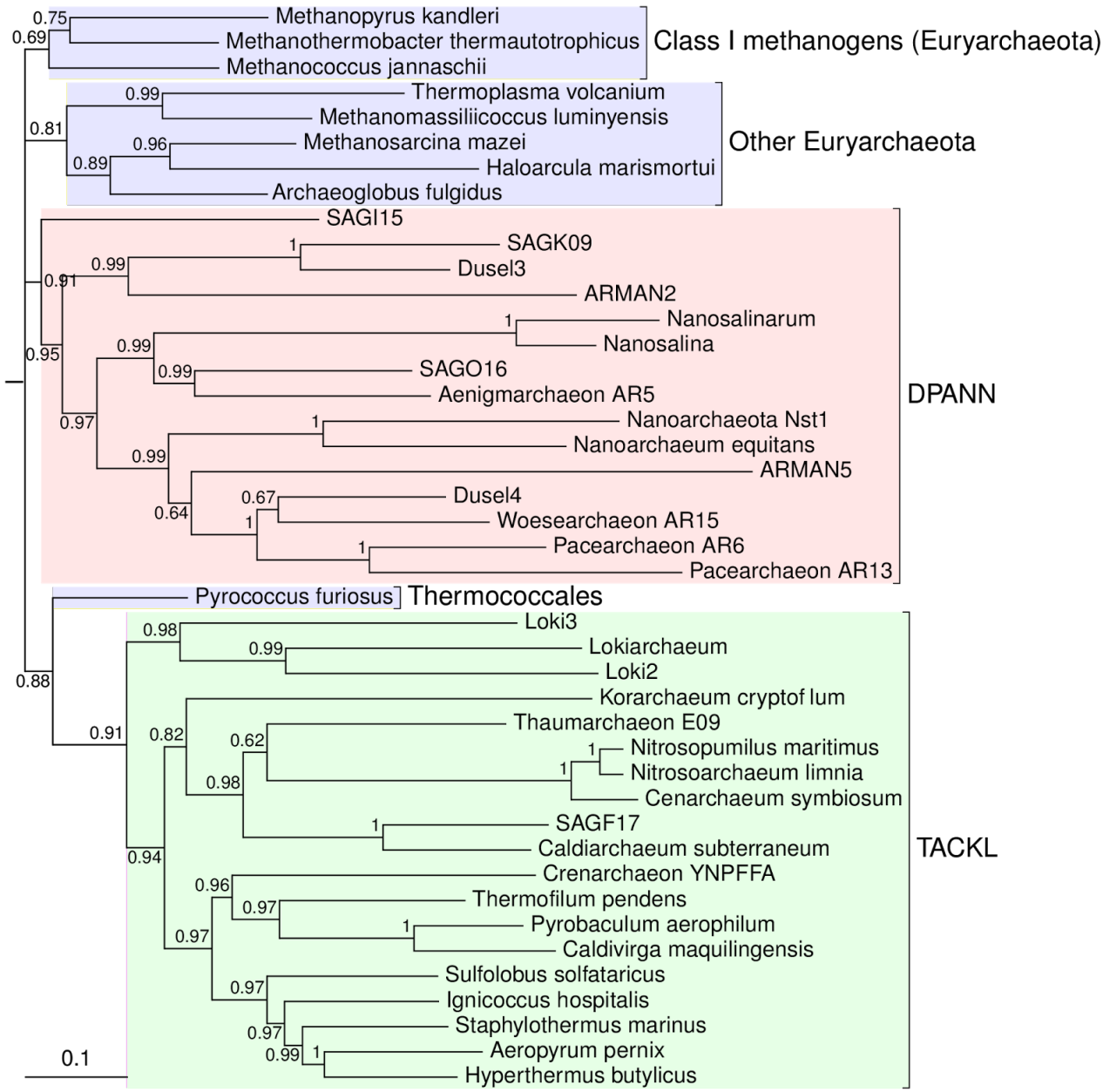


Figure S13: Rooting the Archaea with a bacterial outgroup. This phylogeny was inferred under the CAT+GTR model from a Dayhoff-recoded alignment of 29 broadly-conserved genes in Bacteria and Archaea comprising 8534 aligned amino acid sites; the root nub indicates the point at which the branch leading to Bacteria joins the archaeal in-group. In the Bayesian consensus tree, the root is excluded from the TACKL (PP = 0.91) and DPANN Archaea (PP = 0.91), and from two clades of Euryarchaeota: the class I methanogens (PP = 0.69) and the other Euryarchaeota (PP = 0.81).

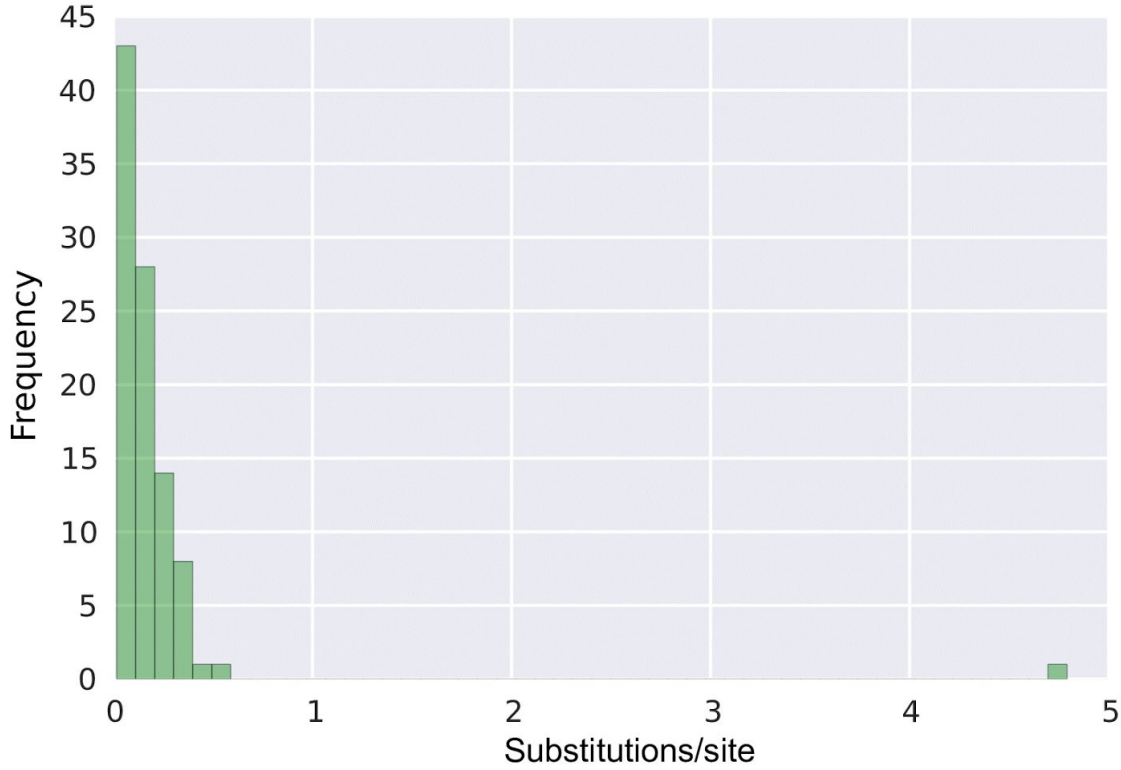


Figure S14: The longest branch leads to the outgroup. In the analysis depicted in Figure S6, the branch separating the bacterial and archaeal clades is by far the longest, with 4.79 expected substitutions/site.

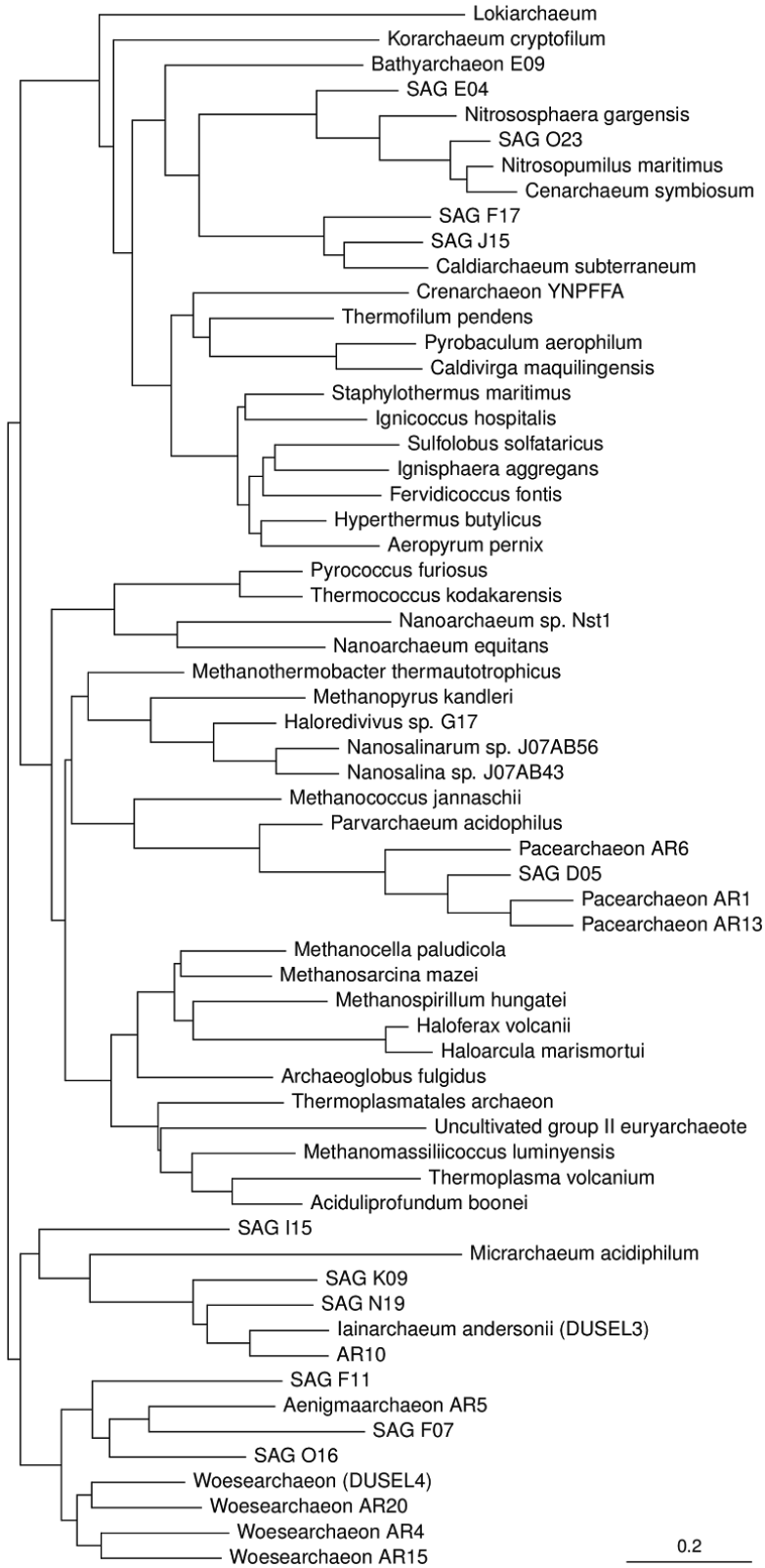


Figure S15: **A candidate tree topology in which DPANN are polyphyletic**, as obtained by combining the results of the single-lineage analyses in an informal supertree. This topology was rejected both by analysis of protein concatenation (PP = 0) and by the DTL model.

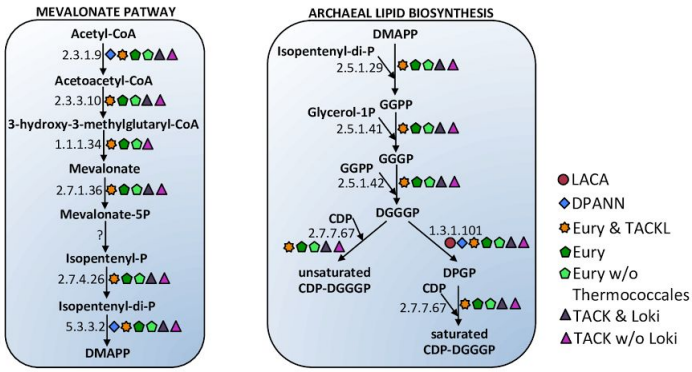


Figure S16: The origins of the archaeal mevaloate and lipid biosynthetic pathways. The last archaeal common ancestor is predicted to have already encoded many of the components of the canonical modern archaeal pathway. An interesting exception is glycerol-1-phosphate dehydrogenase, which was not confidently mapped to LACA due to its absence from a phylogenetically diverse range of archaeal genomes and metagenomes, including those of the group II/III euryarchaeota, some members of the DPANN, and *Lokiarchaeum*. Presence of a gene at a node is indicated by the symbols laid out in the key. Partially filled symbols indicate that only some of the subunits comprising a particular enzyme were present. Abbr.: DMAPP: dimethylallyl diphosphate; GGPP: geranylgeranyl diphosphate; GGGP: geranylgeranylglyceryl phosphate; DGGGP: digeranyl- geranylglyceryl phosphate; CTD-DGGGP: cytidine-diphosphate digeranyl-geranylglyceryl phosphate.

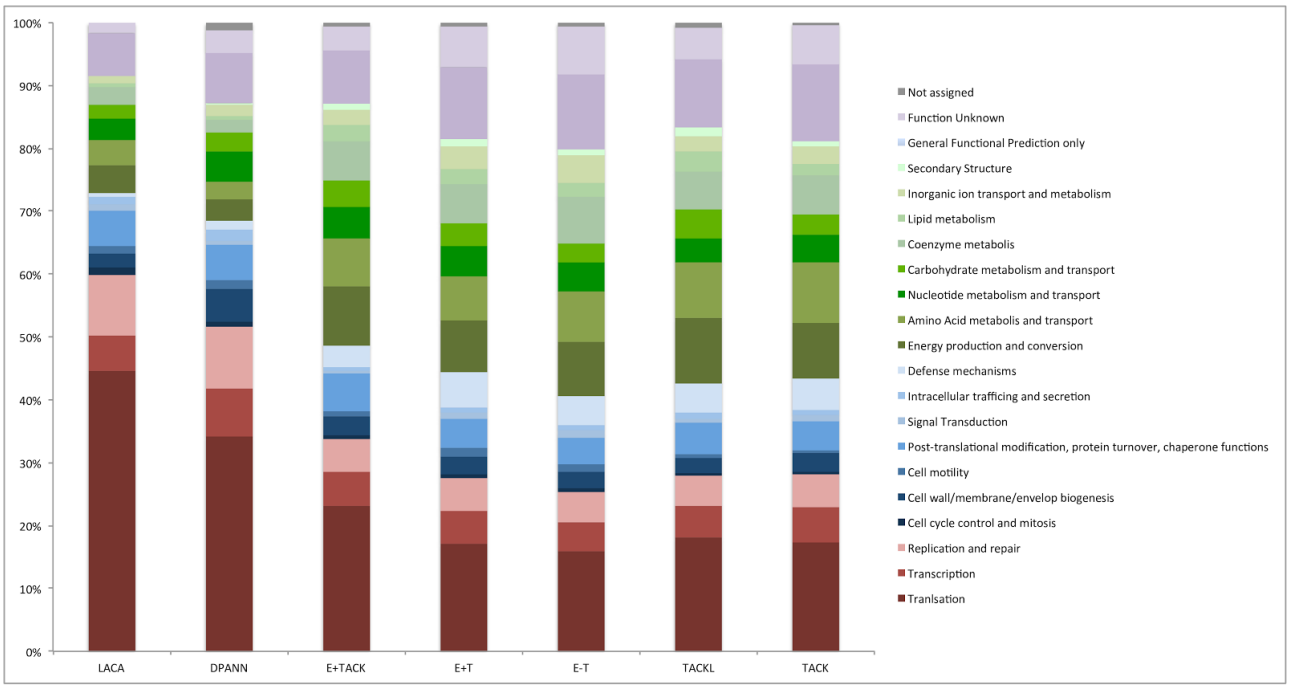


Figure S17: Functional categorization of ancestral gene sets. Bar graph showing the relative abundance of arCOG functional categories assigned to the respective gene repertoire of the analysed ancestors. The total number of proteins that could be assigned to arCOG categories for each of the investigated ancestors was as followed; LACA: 177; DPANN: 337; E+TACK: 660; E+T: 1236; TACKL: 925; TACK: 965. Abbr.: LACA: Last common archaeal ancestor; E: Euryarchaeaota;T: Thermococcales; L: Lokiarchaeota.

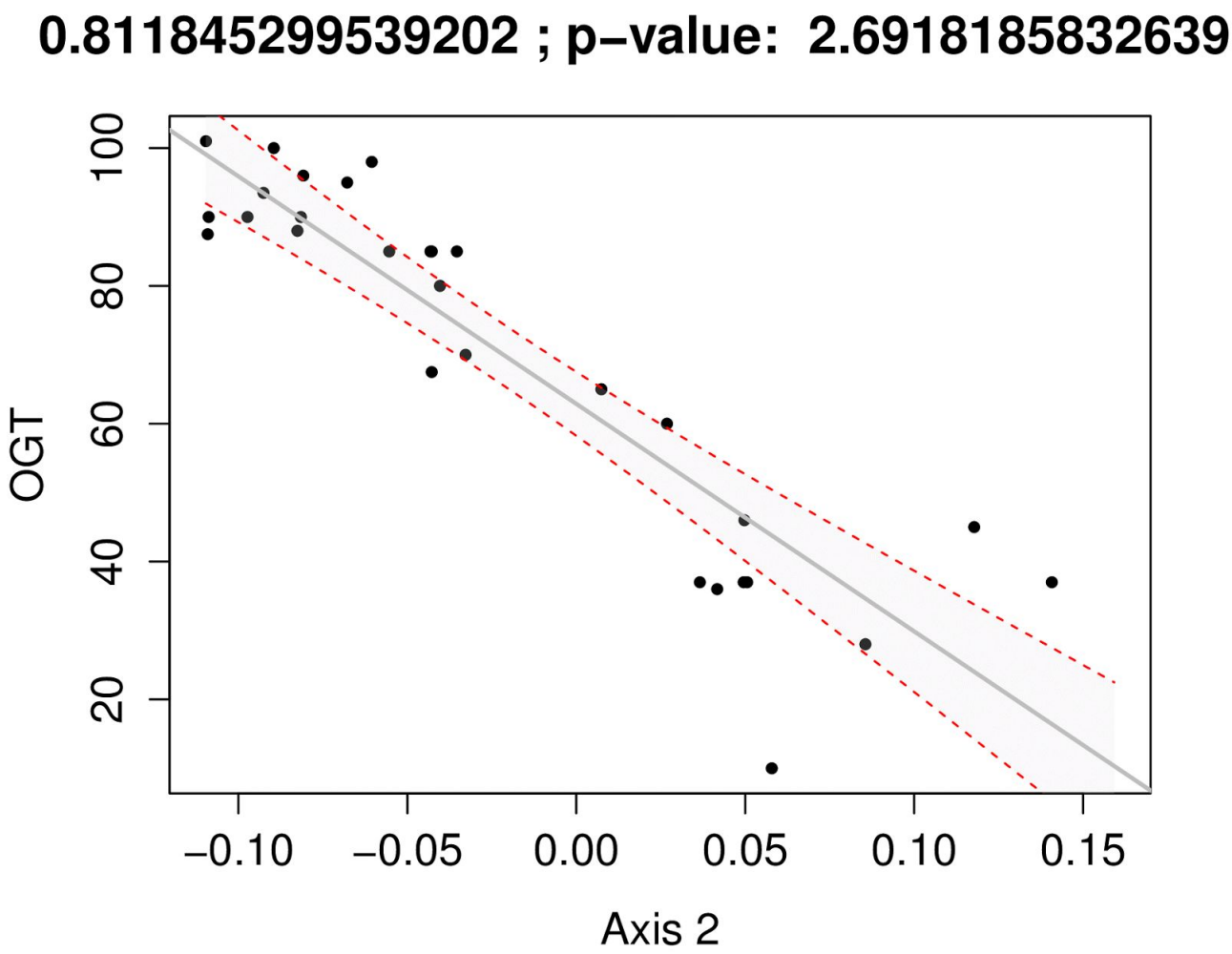


Figure S18: Correlation between optimal growth temperature the second axis of a correspondence analysis on amino acid composition. This correlation was used to predict the growth temperatures of ancestral nodes in the archaeal tree.

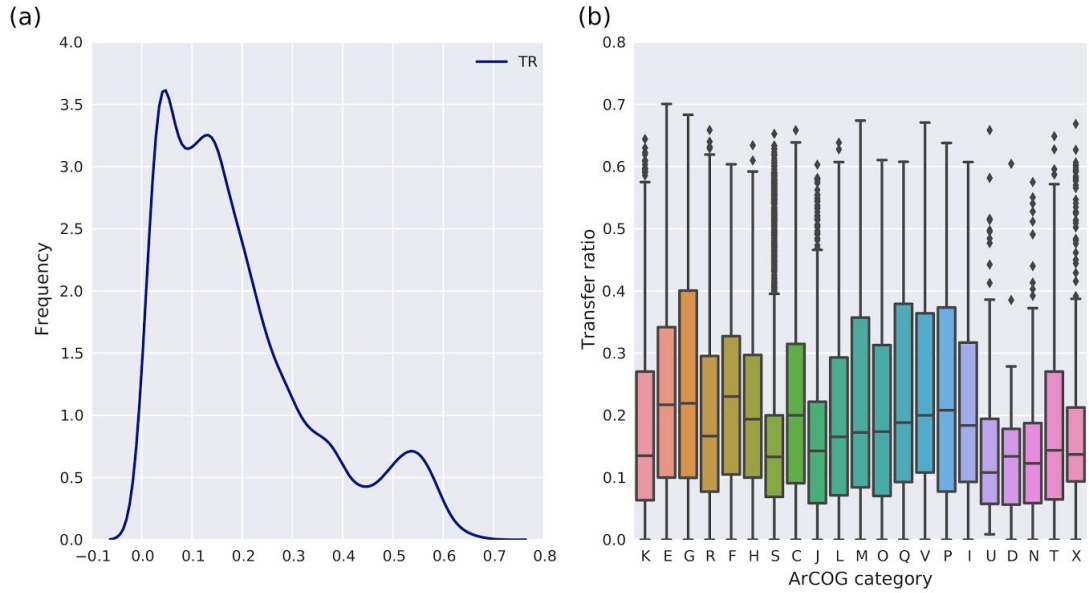


Figure S19: Transferability of archaeal genes. (A) Most gene families experience more vertical than horizontal transmissions (transfer ratio, the proportion of horizontal transfers as a fraction of all transmission events, < 0.5). (B) Transferability varies by gene functional category; genes involved in defense (V) and carbohydrate metabolism (G) are over-represented in the "hump" of genes with TR > 0.5.

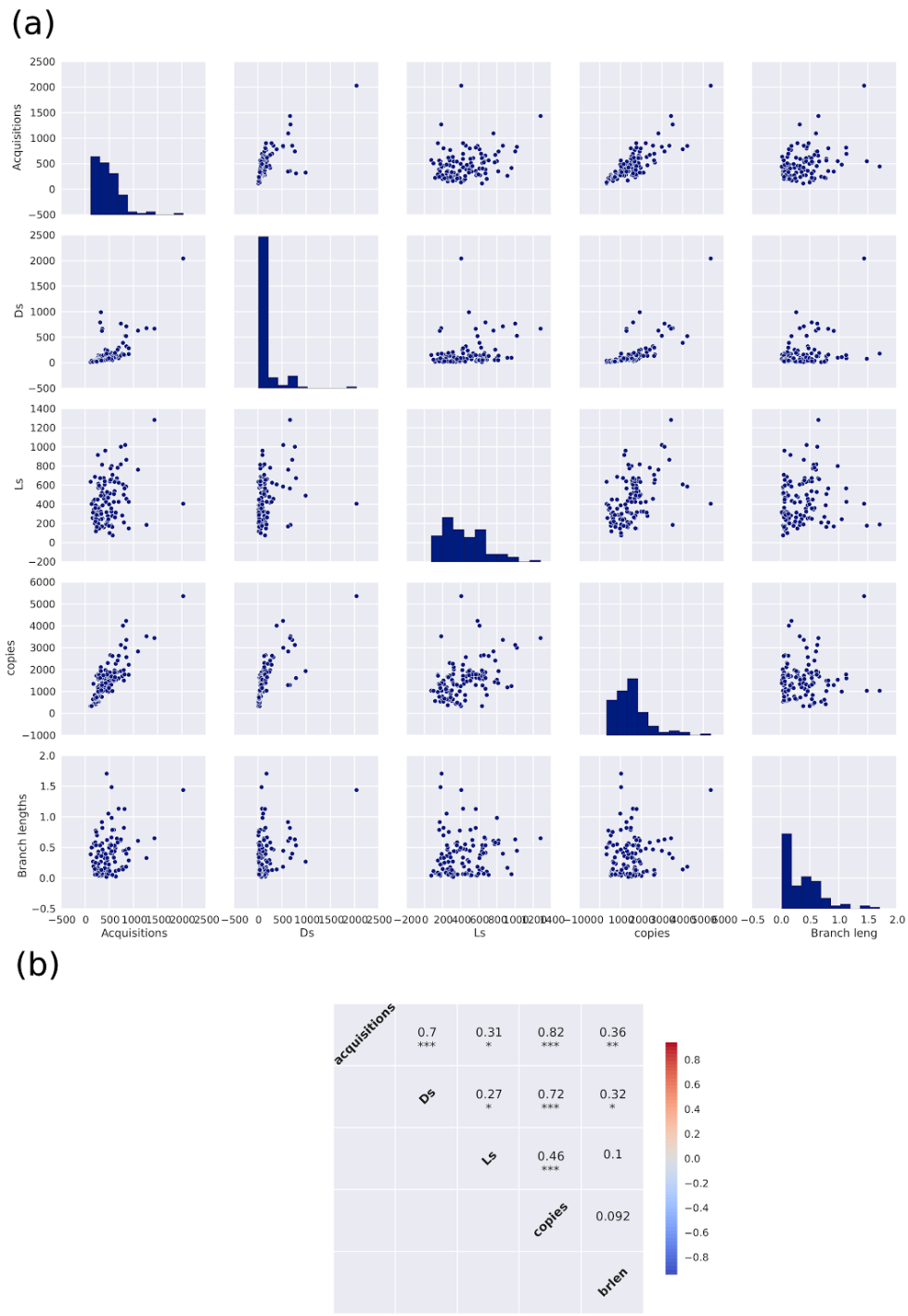


Figure S20: Correlations between rates of genome evolutionary events, proteome size, and concatenated protein branch lengths across the archaeal tree. (A) Distributions and pairwise scatterplots for genome evolutionary events (Gene originations, duplications, transfers, losses, acquisitions (the sum of originations, duplications and transfers), total gene family copy number, and concatenated protein branch lengths on the archaeal tree. (B) Pairwise correlations among these variables.

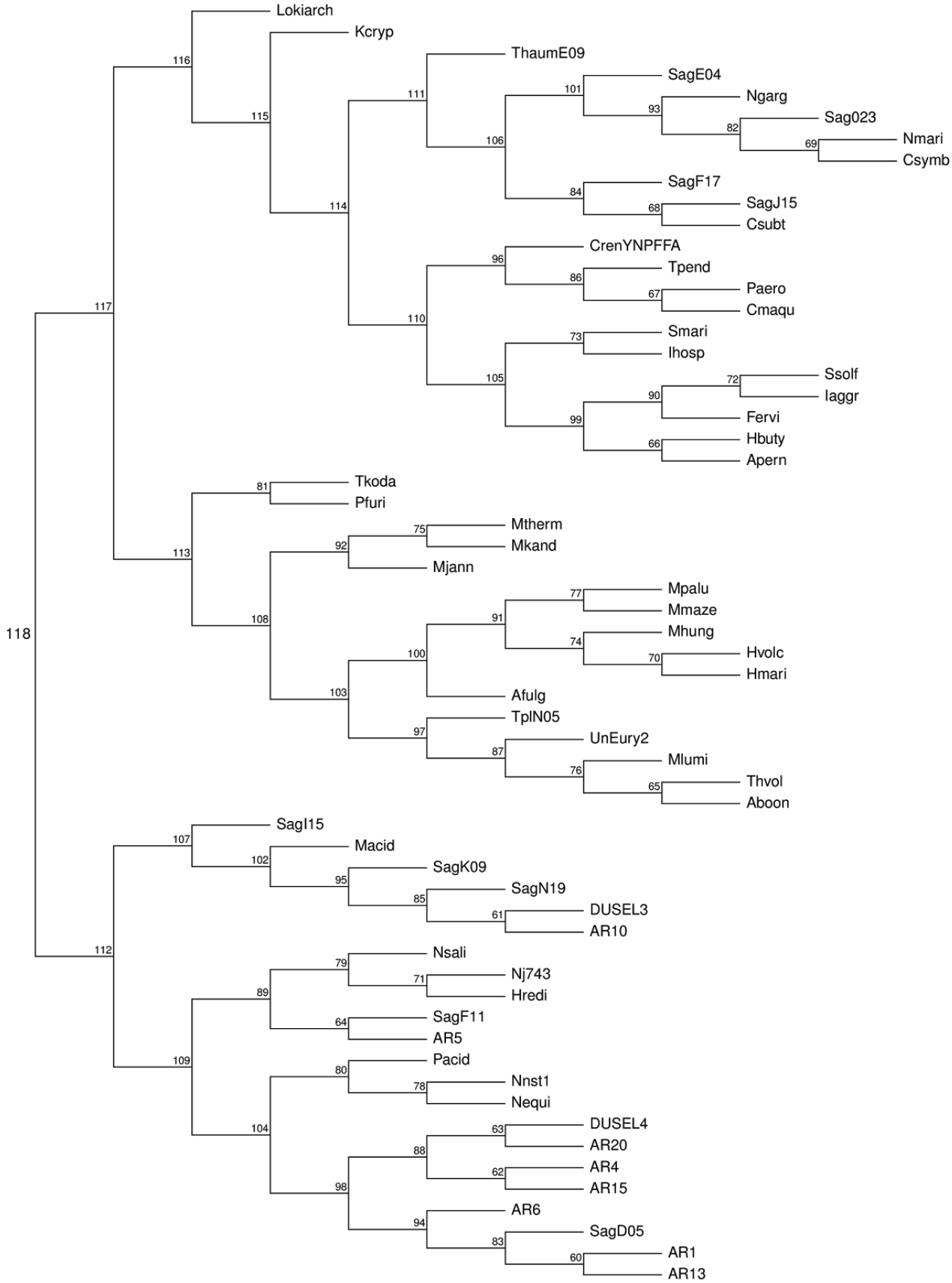


Figure S21: Mapping of internal branch numbers to the rooted archaeal tree. This labelled tree diagram relates the inferred branch-wise numbers of gene originations, duplications, transfers,

losses and total acquisitions (Table S7) to branches on the maximum likelihood rooted archaeal tree.

Supplementary Text

The inferred gene repertoire of the last archaeal common ancestor (LACA). The largest category of genes that can be mapped back to LACA are involved in informational processing machineries such as translation, transcription and replication (Table S6, Figure S9). For instance, many ribosomal proteins, RNA polymerase subunits (e.g. A, B, D and E subunits), tRNA synthetases and the small and large subunits of DNA polymerase II (PolD), proliferating cell nuclear antigen (PCNA) as well as components of the exosome and proteasome were all suggested to have been present in LACA. In addition, at least some subunits of the archaeal secretion system have likely already been part of the proteome of the archaeal ancestor. Beyond these components of the informational processing machinery, a relatively small set of additional genes could be mapped back to LACA (Figure 3, Figure S8).

Notably, and in accordance with previous findings (10–12), the identification of one subunit (cdhC) of the key enzyme of the Wood-Ljungdhal/ Acetyl-CoA pathway in LACA, may suggest that this carbon fixation pathway was encoded by LACA, although other proteins of the Wood-Ljungdahl pathway could not be mapped to LACA with high confidence. In particular, our reconstruction could not resolve whether LACA used inorganic or organic electron donors - hydrogenase subunits were only inferred to the root of Eury and TACKL with sufficient confidence - or whether LACA had the ability to perform methanogenesis. As the Wood-Ljungdahl pathway can function in both autotrophic carbon fixation and heterotrophic growth (13), its presence does not strictly imply that LACA was an autotroph. Other proteins involved in central carbon metabolism included 2-phosphosulfolactate phosphatase (EC 3.1.3.71), that catalyzes the reaction from P-Sulfolactate to Sulfolactate (amongst others functioning in Coenzyme M biosynthesis). This is interesting given the importance of Coenzyme M in the acetyl-CoA pathways as well as methanogenesis.

Furthermore, LACA was inferred to have encoded ribose-phosphate pyrophosphokinase (EC:2.7.6.1), subunit beta of a 2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase, pyruvate, water dikinase (EC:2.7.9.2), hexulose-6-phosphate synthase (EC: 4.1.2.43), AMP phosphorylase (EC:

2.4.2.57/ COG00213) as well as phosphomannomutase/ phosphoglucomutase (EC:5.4.2.8/2).

Evolution of metabolism during the diversification of Archaea (Figure 3, Table S6). Our ancestral reconstructions suggest an anaerobic ancestor of Euryarchaeota and TACKL. The ancestors of these groups were inferred to have encoded a superoxide reductase/ desulfoferredoxin (arCOG02146, pfam01880). This enzyme is common among anaerobic or microaerophilic organisms and catalyzes the detoxification of superoxide to hydrogen peroxide rather than to molecular oxygen, which is the end product of many oxygen-detoxification enzymes present in aerobes (14). In addition, our analyses indicate that the ancestors of TACK and Euryarchaeota encoded F_{420}-dependent, H_2- and/or sulfhydrogenases. This latter enzyme (EC 1.12.98.4/ 1.12.1.3) is comprised of three subunits and functions as both sulfur reductase and hydrogenase in *Pyrococcus furiosus* (15). It is tempting to speculatate that the ancestral enzyme might also have had bifunctionality. NADH dehydrogenases as well as terminal oxidases, in contrast, appeared to have evolved later in aerobic Euryarchaeota and TACK.

Although there was evidence for the early evolution of the Acetyl-CoA pathway, our current investigation could trace back key genes for methanogenesis to the root of Euryarchaeota only. However, the incorporation of genomes from the recently discovered methanogenic or methane-oxidizing Bathyarchaeota (16) in future analyses might help to better clarify the early evolution of methanogenesis and determine whether key genes of this important metabolic pathway could have originated earlier in the archaeal tree.

The bifunctional Fructose-1,6-bisphosphate aldolase/phosphatase (EC: 3.1.3.11), which catalyzes both the synthesis of Fructose-1.6-biphosphate from Glyceraldehyde-3-phosphate and dihydroxyacetone phosphate as well as the subsequent dephosphorylation to Fructose-6-phosphate, was not inferred to any of the ancestral nodes. This was surprising as this protein was suggested to represent an ancestral enzyme of gluconeogenesis (17). This result might reflect the relatively stringent probability threshold we used for ancestral mapping ($P > 0.5$); an alternative possibility is that this gene has experienced horizontal transfer throughout its evolution – for example, the thaumarchaeotal sequence appears derived from Bacteria (17). With the exception of this enzyme, many of the genes of central carbon metabolism, including those involved in glycolysis/gluconeogenesis and the citric acid cycle, could be traced back to the ancestors of

Euryarchaeoa, TACK and Lokiarchaeota, although surprisingly few genes involved in these pathways could be mapped to the root of DPANN. This may, in part, reflect the apparently host-dependent and metabolically reduced lifestyles of known DPANN lineages, although - given our currently limited understanding of DPANN diversity - it seems difficult to reject the possibility that DPANN lineages may encode highly divergent versions of known metabolisms or alternative metabolic pathways that are difficult to assign based on sequence homology alone (18).

Finally, our analyses revealed clear indications for the early evolution of archaeal-type lipids, which are distinct from bacterial lipids and membranes (19). Most proteins involved in the mevalonate pathway and in archaeal lipid biosynthesis have been inferred to the root of Euryarchaeota and TACK, while some components could also be traced back to LACA and the ancestor of DPANN. Additionally, many of the proteins involved in the biosynthesis of amino acids and nucleotides were suggested to have been present in the shared ancestor of TACK and Euryarchaeota but only some of these genes were inferred to have been present at the root of DPANN, while very few of them could be traced back to LACA (Table S6). Our analysis revealed that four subunits of the archaeal-type ATP synthase were inferred to the shared ancestor of TACK and Euryarchaeota and more than six to the subsequent ancestors of Euryarchaeota and TACK, respectively, but only two ATP synthase subunits (B and C) could be confidently traced back to the ancestor of DPANN and none of them to LACA. Na+/or H+-translocating membrane pyrophosphatase (arCOG04949, EC EC 3.6.1.1) as well as an inorganic pyrophasphatase (arCOG01711, EC EC 3.6.1.1) could confidently be traced back to the roots Euryarchaeota and TACK as well as later nodes (Table S6). The absence of these ATP-generating proteins from the inferred gene set of LACA may be a consequence of our conservative mapping approach, in particular because some DPANN Archaea appear to be energy parasites lacking any ATP synthase genes (18, 20, 21). A detailed investigation of the provenance and evolution of these genes within DPANN promises to contribute to our understanding of early membrane bioenergetics (12).

In sum, our findings provide indications that LACA might have been able to fix carbon through the Acetyl-CoA pathway, yet they do not rule out that it had more versatile catabolic and anabolic capabilities. Additionally, this analysis suggests that the metabolic gene complement of archaea has experienced a complex evolutionary history during the diversification of this domain of life, including gene family extinctions and horizontal transfers with other domains, that precludes the

functional attribution of a significant proportion (on the order of 41%, given our modelling assumptions) of the inferred proteome of LACA.

Sensitivity analyses to evaluate the robustness of the ALE method. We performed sensitivity analyses to evaluate the robustness of inferences under ALE to variation in rates of horizontal gene transfer and to biases in taxonomic breadth (species representation) among gene families. To do so, we divided all gene families into quartiles by (i) number of horizontal transfers and (ii) taxonomic breadth, and repeated the rooting analysis (including AU-test of candidate tree topologies) separately for each quartile (Tables S11-12 below). In all of these analyses, a root between DPANN and the other Archaea was the maximum likelihood topology, and the two variants of this topology (in which *Thermoccocales* group either with TACK or Euryarchaeota) were the only trees not rejected by AU-test.

	all	Q1 (0.03-1.51 Transfers)	Q2 (1.49-3.05 Ts)	Q3 (3.05-6.00 Ts)	Q4 (6.00-356.08 Ts)
ΔLL DPANN basal, Th with Eury	**0** **0.66**	**0** **0.945**	**0** **0.99**	**0** **0.163**	**0** **0.583**
ΔLL DPANN basal, Th with TACKL	**12.87023** **0.340**	**11.47753** **0.057**	**19.2287** **0.01**	**-13.3435** **0.837**	**-4.4925** **0.417**
ΔLL *Lokiarchaeum* basal	356.3913 <1e-3	37.78081 <1e-3	89.9036 <1e-3	93.2482 <1e-3	135.4587 <1e-3
ΔLL TACK	406.0855 <1e-3	27.15948 <1e-3	77.0227 <1e-3	87.9706 <1e-3	213.9327 <1e-3
ΔLL TACK+Th	522.1632 <1e-3	63.71319 <1e-3	72.2618 <1e-3	73.7011 <1e-3	312.4871 <1e-3
ΔLL Euryarchaeota basal	614.7883 <1e-3	69.88644 <1e-3	78.8904 <1e-3	86.2237 <1e-3	379.7878 <1e-3
ΔLL Gribaldo et al. root	785.5786 <1e-3	104.1281 <1e-3	119.474 <1e-3	124.6938 <1e-3	437.2827 <1e-3
ΔLL Polyphyletic	4215.381 <1e-3	577.0174 <1e-3	513.9133 <1e-3	728.108 <1e-3	2396.342 <1e-3

DPANN					

Table S11: Tree topology support (difference in likelihood compared to the maximum likelihood topology and AU-test p-value) under the ALE model is robust to variation in levels of horizontal gene transfer. The number of transfers per family was determined by averaging over 100 sampled gene tree reconciliations per family.

	all (4-58 species)	Q1 (4 species)	Q2 (5-6 species)	Q3 (7-13 species)	Q4 (13-58 species)
ΔLL DPANN basal, Th with Eury	**0** **0.66**	**0** **1.0**	**0** **0.99**	**0** **0.595**	**0** **0.105**
ΔLL DPANN basal, Th with TACKL	**12.87023** **0.340**	25.71819 <1e-3	**17.63054** **0.01**	**2.9339** **0.405**	**-33.4124** **0.954**
ΔLL *Lokiarchaeum* basal	356.3913 <1e-3	67.32932 <1e-3	78.10989 <1e-3	121.036 <1e-3	89.9161 <1e-3
ΔLL TACK	406.0855 <1e-3	51.79293 <1e-3	63.66505 <1e-3	103.8843 <1e-3	186.7432 <1e-3
ΔLL TACK+Th	522.1632 <1e-3	51.55903 <1e-3	54.81736 <1e-3	88.1135 <1e-3	327.6733 <1e-3
ΔLL Euryarchaeota basal	614.7883 <1e-3	49.78023 <1e-3	52.39921 <1e-3	109.9062 <1e-3	402.7027 <1e-3
ΔLL Gribaldo et al. root	785.5786 <1e-3	49.98043 <1e-3	82.94379 <1e-3	192.3832 <1e-3	460.2712 <1e-3
ΔLL Polyphyletic DPANN	4215.381 <1e-3	277.625 <1e-3	411.5599 <1e-3	782.9129 <1e-3	2743.283 <1e-3

Table S12: Tree topology support (difference in likelihood compared to the maximum likelihood topology and AU-test p-value) under the ALE model is robust to variation in taxonomic breadth (that is, species representation) among gene families.

Simulation analyses

We also performed analyses to determine whether the ALE_undated method can recover the true root from simulated data. Data were simulated on the following, dated, random 100 species tree:

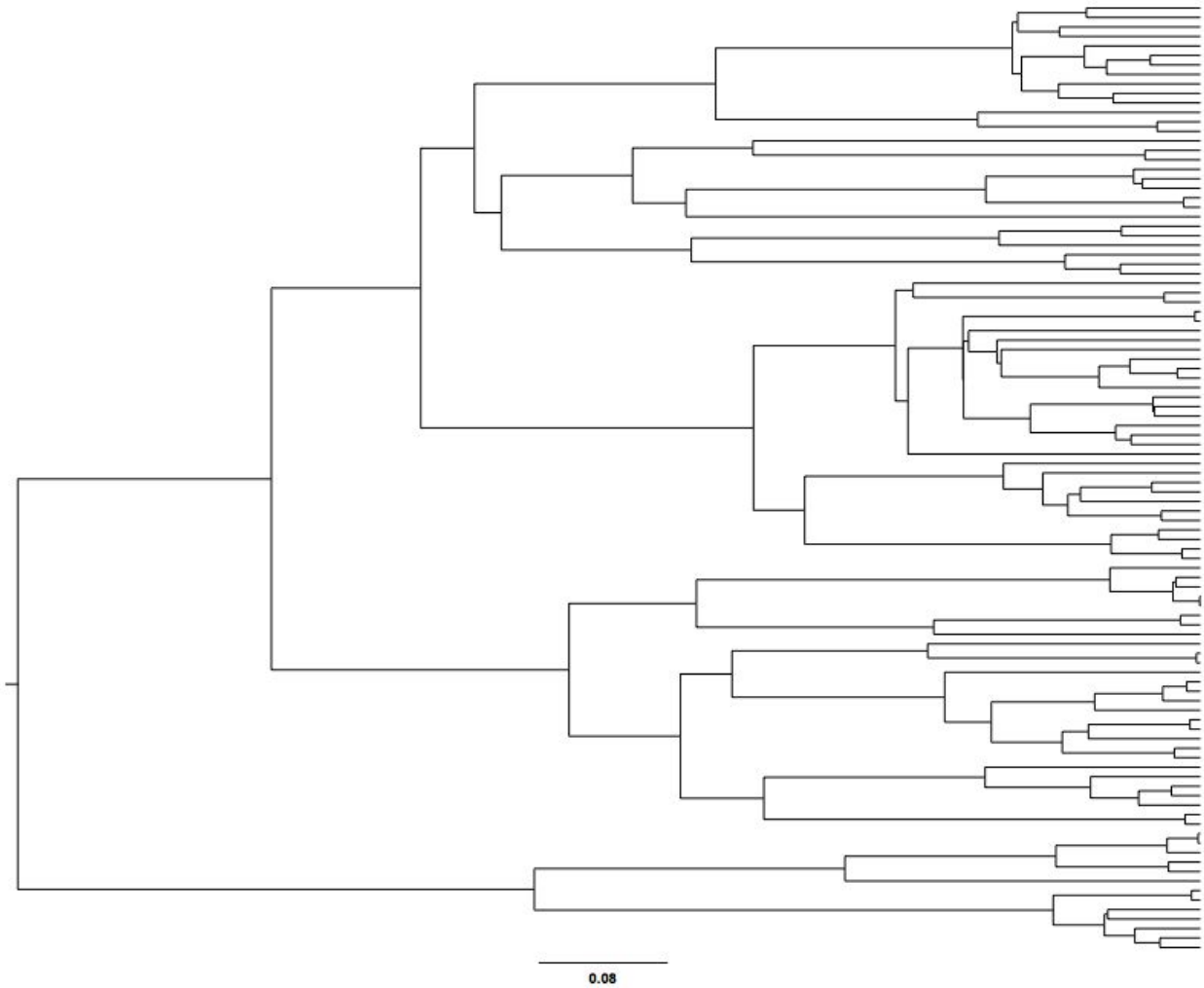


Figure S22: A random 100-species dated tree used in our simulations.

Data were simulated under a more realistic, complex model than that used in the ALE_undated algorithm: a continuous time duplication, transfer, loss, and gene origination (DTLO) process. The branch lengths of the tree, and the values of the D, T, L and O parameters, were tuned empirically to match the real dataset by repeatedly simulating approximately 300 gene families with approximately 10,000 genes.

The tuning was performed as follows: (i) ratios of gene birth (D+T) to death (L) and branch lengths were tuned to match the real dataset - that is, tuning the relative length of internal and external branches while keeping the order of speciations fixed (DT/L ratio from simulation: 0.8528447, DT/L ratio from real data: 0.8715705); (ii) the D and T rates, and the sum of the D, T and L rates, were then individually tuned to match the real data:

T events per gene	Min.	1st Quarter	Median	Mean	3st Quarter	Max
Simulation	0.0	0.0	0.04348	0.07561	0.09545	1.0
Real data	0.0	0.023	0.04	0.10940	0.07000	0.86

D events per gene	Min.	1st Quarter	Median	Mean	3st Quarter	Max
Simulation	0.0	0.0	0.0	0.05501	0.05657	1.0
Real data	0.0	0.0	0.0	0.1062	0.0650	0.9736

A total of 15,543 simulated gene families containing two or more genes, totalling 595,873 simulated genes, were obtained by simulation after fixing the following rate parameters: O = 30, D = 0.5, T = 1.55, L = 2.4, with 200 gene families present at the root. The real dataset has 13,371 families with 2 or more genes, and a total of 85,008 genes. For computational tractability, the following calculations were performed on random subsets of this large simulated dataset.

On a random subset of 20% of these simulated gene families (2500 families, 90,000 genes), we then evaluated the likelihood of each possible root position on the simulation tree using ALEml_undated, the algorithm used in our analyses of real data. The true root was the maximum likelihood root, with $\Delta LL = 92.11$ compared to the next best alternative. Interestingly, ΔLL across all root positions was significantly correlated with the topological distance of the alternative root position (R2=-0.596, *P*=2.2e-16). Thus, our method obtains the true root on simulated data.

To establish a confidence interval for the performance of our method, we ran 100 replicates for the 20 alternative roots topologically closest to the true root on 100 random subsets, each representing 10% of the total simulated dataset. The true root was the maximum likelihood root 95 out of 100 times. In the 5 cases where the true root was not recovered as the ML root, the ML root had a topological distance of 1: that is, it corresponded to one of the four branches closest to to the true root.

Supplementary References

1. Petitjean C, Deschamps P, López-garcía P, Moreira D (2014) Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. 7(1):191–204.
2. Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc Biol Sci* 279(1749):4870–9.
3. Williams TA, Embley TM (2014) Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol Evol* 6(3):474–481.
4. Makarova KS, Wolf YI, Koonin E V (2015) Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel, Switzerland)* 5(1):818–40.
5. Groussin M, Boussau B, Gouy M (2013) A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. *Syst Biol*. 62(4):523-38.
6. Baum BR (1992) Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees. *Taxon* 41(1):3–10.
7. Ragan M a (1992) Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol* 1(1):53–8.
8. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286–8.
9. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:210.
10. Sousa FL, Martin WF (2014) Biochemical fossils of the ancient transition from geoenergetics

to bioenergetics in prokaryotic one carbon compound metabolism. *Biochim Biophys Acta*. 1837(7):964-81.

11. Sousa FL, Nelson-Sathi S, Martin WF (2016) One step beyond a ribosome: the ancient anaerobic core. *Biochim Biophys Acta* 1857(8):1027-38.

12. Lane N, Martin WF (2012) The Origin of Membrane Bioenergetics. *Cell* 151(7):1406–1416.

13. Schuchmann K, Muller V (2016) Energetics and application of heterotrophy in acetogenic bacteria. *Appl Environ Microbiol* 82(14):4056–4069.

14. Jenney Jr. FE, Verhagen MFJM, Cui X, Adams MWW (1999) Anaerobic Microbes: Oxygen Detoxification Without Superoxide Dismutase. *Science* 286(5438):306–309.

15. Ma K, Schicho RN, Kelly RM, Adams MW (1993) Hydrogenase of the hyperthermophile Pyrococcus furiosus is an elemental sulfur reductase or sulfhydrogenase: evidence for a sulfur-reducing hydrogenase ancestor. *Proc Natl Acad Sci U S A* 90(11):5341–5344.

16. Evans PN, et al. (2015) Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* 350(6259):434–8.

17. Say RF, Fuchs G (2010) Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* 464(7291):1077–1081.

18. Castelle CJ, et al. (2015) Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Curr Biol* 25(6):690–701.

19. Sojo V, Pomiankowski A, Lane N (2014) A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol* 12(8):e1001926.

20. Podar M, et al. (2013) Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol Direct* 8(1):9.

21. Mohanty S, et al. (2015) *Structural Basis for a Unique ATP Synthase Core Complex from* Nanoarcheaum equitans *J Biol Chem* 290, 27280-27296.